

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
CENTRO ESTADUAL DE PESQUISAS EM SENSORIAMENTO REMOTO E METEOROLOGIA  
CURSO DE PÓS-GRADUAÇÃO EM SENSORIAMENTO REMOTO

**COMPARAÇÃO DA DISCRIMINAÇÃO LOGÍSTICA  
COM O MÉTODO DA MÁXIMA VEROSSIMILHANÇA GAUSSIANA  
NA CLASSIFICAÇÃO DE IMAGENS DIGITAIS**

Dissertação submetida como  
requisito parcial para a obtenção  
do grau de MESTRE EM  
SENSORIAMENTO REMOTO

HÉLIO RADKE BITTENCOURT

Orientador Professor Dr. Robin Thomas Clarke

Porto Alegre, abril de 2001.

## AGRADECIMENTOS

Gostaria de agradecer especialmente ao Prof. Robin Thomas Clarke, pela orientação prestada durante a elaboração deste trabalho. Desejo também expressar meus agradecimentos ao Prof. Victor Haertel que, juntamente com meu orientador, contribuiu definitivamente para a minha formação acadêmica.

Devido ao excelente convívio proporcionado no decorrer do curso, agradeço também aos colegas José Luis Motta, Mauro Erbert, Marcos Martins e Marcelo Remião; às secretárias Mônica e Magdalena e ao Prof. Jorge Ducati, ex-coordenador do CEPSRM.

Desejo também expressar meus sinceros agradecimentos à CAPES, pelo apoio financeiro, e a Pró-Reitoria de Pesquisa da UFRGS, pelo incentivo à participação em eventos científicos.

Por fim, agradeço a minha esposa Suzel, com quem compartilhei praticamente toda elaboração deste trabalho.

## SUMÁRIO

1	INTRODUÇÃO .....	
1.1	Introdução .....	
1.2	Problema de Pesquisa e Objetivos .....	
2	REVISÃO BIBLIOGRÁFICA .....	
2.1	Histórico do Modelo Logístico e Principais Aplicações .....	
2.2	Discriminação via Modelos Logísticos .....	
2.3	Reconhecimento Estatístico de Padrões .....	
3	MÉTODOS PROPOSTOS E MATERIAL .....	
3.1	Processo de Tomada de Decisão em Reconhecimento de Padrões .....	
3.2	Classificador da Máxima Verossimilhança Gaussiana .....	
3.3	Discriminação Logística .....	
3.4	Material Utilizado .....	
4	RESULTADOS E DISCUSSÕES .....	
4.1	Comparação entre os Processos de Estimação .....	
4.2	Resultados Obtidos por Simulação .....	
4.3	Resultados Obtidos com Imagens Digitais .....	
5	CONCLUSÕES .....	
	REFERÊNCIAS BIBLIOGRÁFICAS .....	

## LISTA DE FIGURAS

FIGURA 1: Função logística e relação <i>logit</i> .....	
FIGURA 2: Número de parâmetros a ser estimado para discriminação de três classes ( $k=3$ ), em função do número de características ( $p$ ) .....	
FIGURA 3: Comportamento das classes $w_1$ , $w_2$ e $w_3$ com relação às características $x_1$ e $x_2$ .....	
FIGURA 4: Saída do procedimento CATMOD do software estatístico SAS .....	
FIGURA 5: Probabilidade condicional $P(w_1   x)$ como função de $x_1$ e $x_2$ .....	
FIGURA 6: Probabilidade condicional $P(w_2   x)$ como função de $x_1$ e $x_2$ .....	
FIGURA 7: Probabilidade condicional $P(w_3   x)$ como função de $x_1$ e $x_2$ .....	
FIGURA 8: Fronteiras de decisão para as classes uniformemente distribuídas ....	
FIGURA 9: Comportamento das classes $w_1$ , $w_2$ e $w_3$ com relação às características $x_1$ e $x_2$ .....	
FIGURA 10: Histogramas bidimensionais ( $x_1$ e $x_2$ ) para as classes $w_1$ , $w_2$ e $w_3$ ...	
FIGURA 11: Saída do procedimento CATMOD do software estatístico SAS .....	
FIGURA 12: Fronteiras de decisão para o modelo de discriminação logística .....	
FIGURA 13: Ilustração das fronteiras de decisão no caso da máxima verossimilhança gaussiana .....	
FIGURA 14: Segmento de uma cena Landsat-TM, composição colorida 5-4-2 (R-G-B) .....	
FIGURA 15: Comportamento espectral médio das classes nas seis bandas espectrais .....	

FIGURA 16: Histogramas da banda TM-3 nas três bandas espectrais .....	
FIGURA 17: Imagem temática gerada a partir da classificação por discriminação logística .....	
FIGURA 18: Imagem temática gerada a partir da classificação por máxima verossimilhança gaussiana .....	
FIGURA 19: Segmento de imagem AVIRIS, composição 96-56-6 (R-G-B) e verdade terrestre .....	
FIGURA 20: Comportamento espectral médio das classes nas dez bandas consideradas.....	
FIGURA 21: Imagens temáticas e verdade terrestre: primeira imagem classificada por máxima verossimilhança gaussiana; segunda imagem classificada por discriminação logística .....	

## LISTA DE TABELAS

TABELA 1: Variáveis <i>dummy</i> $y_i$ para a construção da função de verossimilhança .....
TABELA 2: Exemplo de amostra de treinamento .....
TABELA 3: Vetores média e matrizes covariância por classe .....
TABELA 4: Estimativas para os vetores média e matrizes covariância por classe .....
TABELA 5: Taxa esperada de erros de classificação para as classes uniformemente distribuídas .....
TABELA 6: Tabela de classificação por método .....
TABELA 7: Descrição das classes $w_1$ , $w_2$ e $w_3$ .....
TABELA 8: Tabela de classificação por método baseada em 30.000 padrões .....
TABELA 9: Tabela de classificação baseada em 1.992 amostras de testes .....
TABELA 10: Tabela de classificação de imagem .....

## RESUMO

A tradicional técnica de regressão logística, muito conhecida e utilizada em estudos médicos, permitia apenas a modelagem de variáveis-resposta binárias. A extensão do modelo logístico para variáveis-resposta multinomiais ampliou em muito as áreas de aplicação da regressão logística. Na área de reconhecimento de padrões o modelo logístico multinomial recebeu a denominação de discriminação logística, pois a sua principal função tornou-se classificar/discriminar padrões. A discriminação logística apresenta aparentes vantagens em relação a métodos convencionais de classificação. O método da máxima verossimilhança gaussiana, amplamente difundido e utilizado, necessita da estimação de um número muito grande de parâmetros, pois supõe que as distribuições subjacentes de cada classe sejam normais multivariadas. A discriminação logística, por sua vez, não faz restrições quanto à forma funcional das variáveis, e o número de parâmetros do modelo é relativamente pequeno. Nesse estudo, os princípios da técnica de discriminação logística são apresentados detalhadamente, bem como aplicações práticas de classificação de imagens Landsat-TM e AVIRIS. Os procedimentos de discriminação logística e da máxima verossimilhança gaussiana foram comparados a partir de dados reais e simulados. Os resultados sugerem que a discriminação logística seja considerada como uma alternativa ao método da máxima verossimilhança gaussiana, principalmente quando os dados apresentarem desvios da normalidade.

## ABSTRACT

The traditional Logistic Regression technique, well known and utilized in medical research, allows modeling of binary response variables only. The extension of the Logistic Model to multinomial-response variables significantly broadens the application spectrum of logistic regression. In pattern recognition, the multinomial logistic model has been named Logistic Discrimination, as its main function has become classifying/discriminating patterns. Logistic Discrimination shows apparent advantages in comparison with conventional methods of classification. The Gaussian Maximum Likelihood Method, widely diffused and utilized, requires estimation of a very large number of parameters, for it assumes that the underlying distributions of each class is normal multivariate. Logistic Discrimination, in contrast, has no restrictions with regard to the functional form of the variables, and the number of parameters is relatively small. In this study, the Logistic Discrimination Technique principles as well as practical examples of AVIRIS and Landsat-TM image classification are presented in great detail. We've compared the Logistic Discrimination and the Gaussian Maximum Likelihood procedures using both simulated and real data. Results show that logistic discrimination is an alternative to the Gaussian Maximum Likelihood classifier, specially when the data presents deviations from normality.

# 1 INTRODUÇÃO

## 1.1 *Introdução*

Sensoriamento Remoto pode ser definido, de uma forma geral, como a tecnologia que permite aquisição, armazenamento e interpretação de informações sobre características de objetos de interesse, sem a necessidade de um contato físico com eles. Os objetos de interesse são, na verdade, alvos localizados na superfície terrestre. Os instrumentos que constituem a tecnologia de sensoriamento remoto são câmeras fotográficas, videocâmeras, radares e outros modernos equipamentos sensores instalados a bordo de aeronaves ou satélites.

O produto típico de um sensor remoto é algum tipo de imagem. Nesse estudo foram utilizadas imagens coletadas pelos sistemas sensores eletro-ópticos AVIRIS (aertransportado) e TM (instalado a bordo do satélite Landsat 5). Estes sensores registram os sinais provenientes da interação entre a radiação eletromagnética proveniente do sol e a superfície terrestre na forma de sinais elétricos. Os sinais elétricos, produzidos pelo detector de um sistema eletro-óptico, indicam o nível de radiância médio da área correspondente a um pixel da imagem na

superfície terrestre. Os pixels são os menores elementos de uma imagem, estando diretamente relacionados à resolução espacial dos sistemas sensores.

As imagens coletadas por sensores como o AVIRIS e o TM constituem grandes conjuntos de dados puramente quantitativos, formados por  $n$  observações ( $n = n^{\circ}$  de pixels) e  $p$  variáveis ( $p = n^{\circ}$  de bandas espectrais disponibilizadas pelo sistema sensor). As bandas espectrais são faixas do espectro eletromagnético nas quais o sensor é sensível. Uma mesma cena é imageada por sensores com sensibilidade a diferentes comprimentos de onda, para que assim, seja possível a obtenção de um conjunto de características para um mesmo alvo.

De um ponto de vista matemático, cada pixel da imagem pode ser visualizado como um ponto num espaço  $p$ -dimensional. Se admitirmos que a radiância dos alvos captada pelo sensor é influenciada por inúmeros fatores aleatórios (por exemplo presença da atmosfera, nível de iluminação solar, deficiências no próprio sistema sensor), podemos considerar o vetor de valores observados para cada pixel como uma estimativa do verdadeiro nível de radiância do(s) alvo(s), correspondente(s) àquele pixel. Pensando desta forma, ou seja, incluindo um componente de incerteza aos valores observados para cada pixel, passaremos a ter um problema estatístico quando desejarmos fazer uma classificação da imagem e, portanto, modelos probabilísticos deverão ser preferencialmente utilizados.

Tratando-se de problemas estatísticos no espaço multidimensional, as técnicas de análise estatística multivariada vêm a se enquadrar perfeitamente. A palavra *multivariada* é utilizada quando estamos trabalhando num espaço de três ou mais dimensões. Em uma imagem Landsat-TM o número de bandas espectrais disponibilizadas pelo sistema sensor é sete e, portanto, podemos trabalhar num espaço de sete dimensões ( $p=7$ ). Já o sensor AVIRIS, classificado como um sensor hiperespectral, disponibiliza 220 bandas espectrais ( $p=220$ ).

Problemas de classificação de imagens obtidas por meio de sensores remotos também se enquadram numa importante área, denominada *Reconhecimento de Padrões*, que trabalha com a identificação, descrição, agrupamento e classificação de padrões. Exemplos de padrões podem ser letras de um alfabeto, árvores de uma floresta e pixels de uma imagem. A área de reconhecimento de padrões é de fundamental importância no mundo moderno, onde cada vez mais necessitamos de sistemas, preferencialmente automáticos, que sejam capazes de lidar com a grande quantidade de informação disponível.

Os problemas estatísticos de classificação de imagens digitais obtidas por sensoriamento remoto são estudados de forma mais aprofundada na área de *Reconhecimento Estatístico de Padrões*, onde é feita a integração entre técnicas estatísticas e vários tipos de problemas envolvendo reconhecimento de padrões.

Existem muitos classificadores já consagrados e amplamente utilizados, como, por exemplo, o da máxima verossimilhança gaussiana. Esse classificador pode

ser encontrado em praticamente todos pacotes computacionais disponíveis no mercado. Contudo, existem situações onde o classificador baseado na distribuição gaussiana pode não produzir os melhores resultados, pois a suposição de normalidade nem sempre pode ser aceita.

Nesse estudo será apresentado um classificador baseado em um conceito ainda pouco explorado na área de reconhecimento de padrões – a discriminação logística – e sua comparação com o método da máxima verossimilhança gaussiana. A primeira vista, a discriminação logística apresenta duas vantagens diretas em relação ao método da máxima verossimilhança gaussiana: 1) não é necessário fazer suposições quanto à forma funcional dos dados de entrada; 2) o número de parâmetros a ser estimado é relativamente pequeno.

A técnica de regressão logística é amplamente difundida na área médica, principalmente em problemas que envolvam variáveis-resposta binárias. Contudo ainda é pouco explorada do ponto de vista discriminatório para classificação de imagens digitais. Portanto, no decorrer desse trabalho, pretendemos apresentar a técnica de discriminação logística como uma alternativa a métodos convencionais de classificação e, principalmente, compará-la com o método da máxima verossimilhança gaussiana.

## *1.2 Problemas de Pesquisa e Objetivos*

A nova geração de sensores hiperspectrais tem capacidade de disponibilizar informação numa grande quantidade de bandas espectrais. O sensor

aerotransportado AVIRIS pode imagear uma mesma cena em 220 faixas do espectro eletromagnético, enquanto sensores como o TM, instalado no satélite Landsat-5, e o HRV, instalado no satélite SPOT, ficam limitados a, no máximo, sete bandas.

A grande quantidade de informação presente em uma imagem AVIRIS proporciona riqueza de detalhes, contudo o processamento da imagem torna-se mais complicado. Em procedimentos de reconhecimento estatístico de padrões que utilizam métodos paramétricos como, por exemplo, a discriminação logística e o classificador gaussiano, um aumento no número de bandas espectrais pode causar um aumento tão grande no número de parâmetros a ponto de comprometer o processo de estimação.

A discriminação logística apresenta aspectos altamente desejáveis em classificadores paramétricos. Foram exatamente esses fatores que motivaram a execução desse estudo, cujos principais objetivos são listados a seguir na forma de perguntas:

- Qual a eficiência do modelo logístico quando os dados nas  $p$  bandas espectrais de uma imagem digital têm distribuição normal multivariada?
- Qual o comportamento da discriminação logística frente a dados claramente não gaussianos?
- Qual a eficiência da discriminação logística e do método da máxima verossimilhança gaussiana na classificação de imagens Landsat-TM e AVIRIS?

## 2 REVISÃO BIBLIOGRÁFICA

De acordo com Hosmer e Lemeshow (1989, p. 6) a regressão logística, em sua forma tradicional, consiste de um modelo que relaciona um conjunto de  $p$  variáveis independentes  $X_1, X_2, \dots, X_p$  a uma variável dependente  $Y$  que assume apenas dois possíveis estados, digamos 0 ou 1. O modelo logístico permite a estimação direta da probabilidade de ocorrência de um evento ( $Y=1$ ):

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$$

e, conseqüentemente,

$$P(Y = 0) = 1 - P(Y = 1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$$

onde  $\beta_i$  são os parâmetros do modelo.

A transformação que está por trás do modelo logístico é a chamada transformação *logit*, denotada por  $g(x)$ . É uma função linear nos parâmetros  $\beta_i$ , contínua e que pode variar de  $-\infty$  a  $+\infty$ :

$$\text{logit}(x) = g(x) = \ln \left[ \frac{P(Y=1)}{1 - P(Y=1)} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Hosmer e Lemeshow (1989, p. 6) dizem que há pelo menos duas razões para utilização do modelo logístico na análise de variáveis-resposta dicotômicas: 1) de um ponto de vista matemático, é extremamente flexível e fácil de ser utilizada; 2) permite uma interpretação de resultados bastante rica e direta. A figura 1 apresenta a função logística com o seu característico formato em ‘S’ e a relação linear entre uma única variável  $x$  e o *logit*  $g(x)$ .

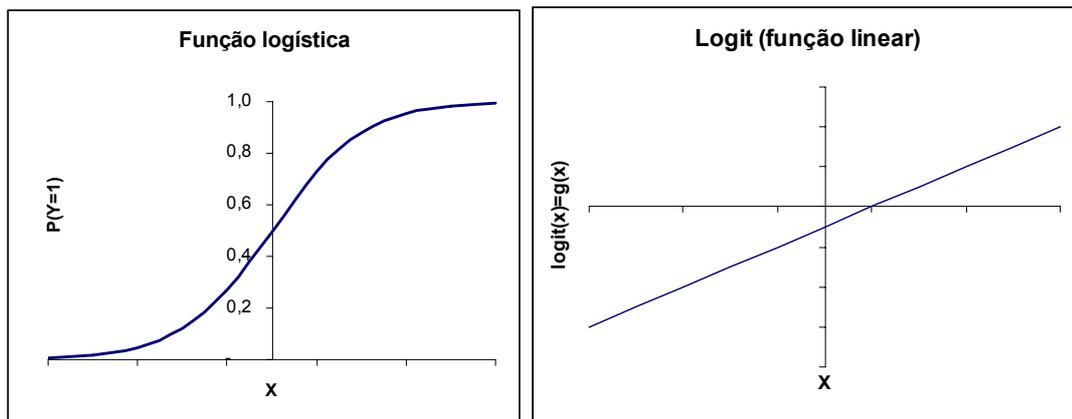


FIGURA 1: Função logística e relação *logit*

Anderson (1982) enfatiza que a discriminação logística pode ser aplicada a uma grande variedade de famílias de distribuições, pois a suposição de linearidade do *logit* é válida numa grande quantidade de funções de distribuição de probabilidade, como por exemplo, a distribuição normal multivariada com igual matriz covariância.

## 2.1 Histórico do Modelo Logístico e Principais Aplicações

É difícil precisar exatamente o ano no qual o modelo logístico foi utilizado pela primeira vez, mas Cox e Snell (1989, p. 25) e Hosmer e Lemeshow (1989) concordam que o modelo de regressão logística ganhou reconhecimento após o trabalho de Truett, Cornfield e Kennel (1967) que analisava o risco de doença coronária em um grande projeto conhecido por “*Framingham heart study*”. Esse trabalho ganhou fama e até hoje é considerado um marco inicial dos estudos envolvendo regressão logística nas áreas da saúde. McLachlan (1992, p. 256) também afirma que as primeiras aplicações do modelo logístico foram em estudos prospectivos de doenças coronárias. Contudo, nessas aplicações, os autores realizaram o processo de estimação de parâmetros sob a suposição de normalidade, que se torna desnecessária quando a estimação é feita por máxima verossimilhança via métodos numéricos. O procedimento de estimação em um contexto mais genérico foi proposto por Day and Kerridge (1967) e por Walker and Duncan (1967).

A importância da distribuição logística e as vantagens das estimativas obtidas com a transformação *logit* já eram conhecidas nas décadas de 40 e 50 para análise de variáveis-resposta binárias. Berkson apud Cox e Snell (1989, p. 24) lançou uma série de artigos onde apresentava o modelo logístico como uma alternativa para a análise de dados binários e também como sendo preferível ao modelo *probit*<sup>1</sup>. McLachlan (1992, p. 257) afirma que os modelos *logit* e *probit* são praticamente

---

<sup>1</sup>  $probit(x) = \Phi^{-1}[P(Y = 1)]$  onde  $\Phi [ \cdot ]$  denota a probabilidade acumulada na normal padrão.

indistinguíveis na prática e que o modelo logístico é preferido por sua conveniência computacional.

Hosmer e Lemeshow (1989) afirmam que o modelo de regressão logística tornou-se um método padrão de análise de regressão de dados dicotômicos, especialmente nas ciências da saúde. De fato, aplicações da regressão logística são comumente encontradas em periódicos da área de saúde, tais como *The American Journal of Epidemiology*, *The American Journal of Public Health*, *The International Journal of Epidemiology* e *The Journal of Chronic Diseases*.

A literatura sobre regressão logística é muito vasta, tendo apresentado um crescimento muito rápido. Além das inúmeras aplicações na área da saúde, a regressão logística também tem sido utilizada no campo da econometria, administração e educação. Por esse motivo, encontramos artigos envolvendo regressão logística em periódicos de diversas áreas.

## *2.2 Discriminação via Modelos Logísticos*

Segundo McLachlan (1992, p. 257) o modelo logístico foi primeiramente introduzido sem a preocupação de utilizá-lo para discriminação de classes. É comum encontrarmos na literatura os termos *regressão logística* e *discriminação logística*. Ambos praticamente se referem à mesma coisa, mudando, às vezes, apenas o ponto de vista. Quando o principal objetivo da utilização do modelo logístico consiste na discriminação/classificação de elementos, o termo discriminação logística é o mais

indicado. Nas áreas relacionadas à saúde, onde o modelo logístico é muito utilizado para relacionar uma variável binária com um conjunto de variáveis independentes, o termo mais empregado é regressão logística.

A discriminação logística só tornou-se possível porque o modelo, originalmente desenvolvido para variáveis-resposta binárias, foi estendido para variáveis-resposta politômicas (três ou mais categorias). Hosmer e Lemeshow (1989, p. 216-245) apresentam a extensão do modelo logístico para variáveis politômicas utilizando como exemplo uma variável resposta  $Y$  com três categorias, codificadas como 0, 1 ou 2.

Assim como a função *logit* de uma variável resposta binária foi parametrizada em termos da razão entre a probabilidade de ocorrência ( $Y=1$ ) e não-ocorrência de um evento ( $Y=0$ ), o modelo logístico para uma variável resposta com três níveis terá duas funções *logit*: a razão entre  $Y=1$  e  $Y=0$  e a razão entre  $Y=2$  e  $Y=0$ . Nesse caso, o nível  $Y=0$  é assumido como base.

$$g_1(x) = \ln \left[ \frac{P(Y=1)}{P(Y=0)} \right] = \beta_{10} + \beta_{11}x_1 + \dots + \beta_{1p}x_p$$

$$g_2(x) = \ln \left[ \frac{P(Y=2)}{P(Y=0)} \right] = \beta_{20} + \beta_{21}x_1 + \dots + \beta_{2p}x_p$$

A partir das funções lineares  $g_i(x)$ , cujos parâmetros são estimados por máxima verossimilhança, é possível calcular as probabilidades condicionais de

ocorrência de cada categoria da variável-resposta  $Y$  dado um conjunto de observações  $x_1, x_2, \dots, x_p$ , conforme segue:

$$P(Y = 0 | x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$P(Y = 1 | x) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$P(Y = 2 | x) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

Krzanowski (1988, p. 359) afirma que, no momento em que as probabilidades *a posteriori*  $P(Y=y|x)$  do modelo logístico são utilizadas para se estabelecer uma regra de alocação, a abordagem é chamada de *discriminação logística*. Contudo trabalhos mais antigos, como o de Press e Wilson (1978) e Efron (1975), utilizavam o termo *regressão logística* mesmo quando o modelo era utilizado para discriminação.

A extensão do modelo para  $k$  categorias é direta e será apresentada em detalhes no próximo capítulo, bem como os procedimentos de estimação por maximização da função de verossimilhança.

### 2.3 Reconhecimento Estatístico de Padrões

Em recente revisão sobre o estado-da-arte na área de reconhecimento de padrões, Jain, Duin e Mao (2000) fizeram um rápido levantamento nas edições da

revista *IEEE Transactions on Pattern Analysis and Machine Intelligence*, onde verificaram que, desde 1979, 350 artigos da revista enquadravam-se na área de reconhecimento de padrões. A esmagadora maioria dos artigos (86%) relacionava-se à abordagem estatística, sendo que mais da metade deles tratava do delineamento de classificadores.

Segundo Jain, Duin e Mao (2000, p. 4), a abordagem estatística tem sido estudada de forma muito intensa, sendo atualmente a mais utilizada na prática, apesar das técnicas envolvendo redes neurais estarem em ampla expansão.

Na abordagem estatística cada padrão (em imagens digitais um padrão é um pixel)  $x_i$  é representado em termos de  $p$  características, sendo encarado como um vetor no espaço  $p$ -dimensional, onde  $p$  pode ser, por exemplo, o número de bandas espectrais. O principal objetivo é a definição de fronteiras no espaço  $p$ -dimensional que determinem a separação entre diferentes classes de padrões presentes na imagem. As fronteiras de decisão são determinadas por distribuições probabilísticas que definem a probabilidade dos padrões pertencerem a cada classe, onde o principal objetivo é a minimização da taxa de erro de classificação.

Processos de reconhecimento de padrões que utilizam a abordagem estatística exigem que o pesquisador faça algumas escolhas durante a sua realização. A primeira escolha define o método de classificação que será utilizado: classificação supervisionada ou não-supervisionada. Em métodos supervisionados, padrões são identificados como membros de classes pré-definidas. Quando temos amostras de padrões representativas das classes presentes na imagem (amostra de treinamento), as

técnicas supervisionadas são preferíveis. Em muitas situações, a escassez de amostras pode impossibilitar a realização de processos supervisionados. Nesses casos, os padrões podem ser agrupados de acordo com similaridades existentes entre eles, sem a necessidade de amostras de treinamento, o que caracteriza os processos não-supervisionados. O termo análise de agrupamento (*cluster analysis*) é utilizado quase como um sinônimo desse tipo de classificação.

A segunda importante escolha que o pesquisador deve fazer consiste da opção entre métodos paramétricos e não-paramétricos. Se a forma das funções de probabilidade dentro de cada classe é conhecida, então um conjunto de parâmetros deverá ser estimado. No caso da distribuição normal multivariada, o vetor média e a matriz covariância são os parâmetros que definem a distribuição. Na discriminação logística, os vetores de parâmetros  $\beta_i$  definem as funções *logit*. Os métodos não-paramétricos, por sua vez, são utilizados quando a forma da função de probabilidade das classes presentes na imagem não é conhecida. Nesses casos, não é necessária a especificação de um modelo que descreva o comportamento probabilístico dos padrões dentro de cada classe. Por esse motivo, métodos não-paramétricos também são chamados de métodos de distribuição livre.

Tanto o tradicional método da máxima verossimilhança gaussiana quanto a discriminação logística são métodos supervisionados e paramétricos, pois necessitam de um conjunto de amostras de treinamento a partir do qual os parâmetros necessários são estimados. McLachlan (1992, p. 255) prefere denominar a discriminação logística como uma abordagem parcialmente paramétrica, pois são apenas as funções *logit* que são assumidas ser lineares.

### 3 MÉTODOS PROPOSTOS E MATERIAL

#### *3.1 Processo de Tomada de Decisão em Reconhecimento de Padrões*

A abordagem estatística para reconhecimento de padrões está intimamente relacionada com teorias de decisão estatística e áreas da análise multivariada. Nesse tipo de abordagem, cada padrão é considerado como uma entidade única, representada por um número finito de características, dispostas na forma de um vetor. No caso específico de imagens digitais, cada pixel da imagem constitui um vetor, formado por um conjunto de  $p$  características, que podem ser, por exemplo, espectrais ou texturais.

$$\text{Padrão } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

Vamos considerar o problema de classificação de um dado padrão  $\mathbf{x}$  em uma das  $k$  possíveis classes,  $w_1, w_2, \dots, w_k$ , partindo do pressuposto que cada padrão pode ser alocado em uma, e somente uma classe. A determinação de regras para alocação de padrões em classes pré-definidas pode ser vista como um problema de

análise discriminante. A quantidade e a qualidade da informação disponível sobre os padrões é que irão possibilitar a determinação de regras mais ou menos eficientes para a correta discriminação dos padrões em suas respectivas classes.

Tou e Gonzalez (1974, p. 111) consideram que o processo de tomada de decisão em reconhecimento de padrões, pode ser tratado como uma disputa entre o classificador do sistema de reconhecimento de padrões e a natureza. Nesse jogo, o objetivo do classificador é encontrar uma solução ótima a qual minimize a probabilidade de erros de classificação.

A regra de decisão de Bayes é um dos possíveis caminhos para se encontrar a solução ótima. Essa estratégia parte do pressuposto que as funções de probabilidade dentro de cada classe  $w_i$ , denotas por  $p(\mathbf{x} | w_i)$ , são conhecidas e que há disponibilidade de informações adicionais sobre as classes, as quais chamaremos de probabilidades *a priori*  $P(w_i)$ . Portanto, a partir do conhecimento das funções de probabilidade de cada classe e das probabilidades *a priori*, é possível utilizar o bem-conhecido teorema de Bayes para estabelecer a probabilidade *à posteriori* de um certo padrão  $\mathbf{x}$  pertencer a classe  $w_i$ .

$$\text{Teorema de Bayes: } P(w_i | \mathbf{x}) = \frac{P(w_i) \times p(\mathbf{x} | w_i)}{\sum_{j=1}^k P(w_j) \times p(\mathbf{x} | w_j)}$$

onde,

$P(w_i)$  = probabilidade *a priori* de um padrão pertencer à classe  $w_i$ .

$k$  = número de classes

$p(\mathbf{x} | w_i)$  = função densidade de probabilidade de  $\mathbf{x}$  na classe  $w_i$

$P(w_i | \mathbf{x})$  = probabilidade *a posteriori* do padrão  $\mathbf{x}$  pertencer à classe  $w_i$

Logicamente  $\sum_{i=1}^k P(w_i | \mathbf{x}) = 1$ . Como o denominador do teorema de

Bayes é constante para todas as classes  $w_i$ , ele pode ser desconsiderado na regra de classificação.

A partir daí podemos estabelecer uma simples regra de classificação baseada apenas no numerador do teorema de Bayes. Assim, o padrão  $\mathbf{x}$  será classificado como pertencente à classe  $w_i$  se:

$$P(w_i | \mathbf{x}) > P(w_j | \mathbf{x}), \text{ para todo } j \neq i \quad \text{ou similarmente}$$

$$P(w_i) \times p(\mathbf{x} | w_i) > P(w_j) \times p(\mathbf{x} | w_j) \text{ para todo } j \neq i.$$

O custo de classificar erroneamente um dado padrão  $\mathbf{x}$  na classe  $w_i$ , quando na realidade o mesmo deveria ser classificado na classe  $w_j$  ( $i \neq j$ ), pode ser utilizado para estabelecer uma regra de classificação baseada na minimização da função risco  $R(w_i | \mathbf{x})$ :

$$R(w_i | \mathbf{x}) = \sum_{j=1}^k L(w_i, w_j) \times P(w_j) \times p(\mathbf{x} | w_j)$$

onde,  $L(w_i, w_j)$  é o custo de alocar o padrão  $\mathbf{x}$  na classe  $w_i$  quando a classe verdadeira é  $w_j$ . Nesse caso, um dado padrão  $\mathbf{x}$  será classificado como pertencente à classe  $w_i$  na qual o risco  $R(w_i | \mathbf{x})$  é mínimo.

### 3.2 Classificador da Máxima Verossimilhança Gaussiana

Quando há conhecimento de que as funções densidade de probabilidade  $p(\mathbf{x} | w_i)$  nas classes  $w_i$  ( $i=1,2,\dots,k$ ) sejam gaussianas, ou haja evidências suficientes para aceitação de tal suposição, a regra de classificação de Bayes toma uma forma bastante conhecida e utilizada. McLachlan (1992, p. 52) e Tou e Gonzalez (1974, p. 119) afirmam que modelos normais para as funções densidade nas  $k$  classes são importantes tanto do ponto de vista teórico como prático, sendo apropriados para muitas importantes aplicações práticas.

Johnson e Wichern (1982, p. 124) salientam que, apesar de dados reais nunca serem exatamente normais, a função densidade gaussiana é frequentemente uma aproximação útil para a verdadeira distribuição populacional. Muitos problemas do mundo real inserem-se naturalmente dentro da estrutura da teoria normal. Haertel e Landgrebe (1999, p. 2374) informam que, tratando-se de cenas naturais, as distribuições das classes espectrais presentes na imagem podem ser aproximadas pela distribuição normal multivariada.

A bem-conhecida distribuição normal  $p$ -variada é dada por

$$p(\mathbf{x} | w_i) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)' \Sigma_i^{-1}(\mathbf{x} - \mu_i)\right\} \quad i = 1, 2, \dots, k$$

onde  $\mu_i$  e  $\Sigma_i$  são, respectivamente, os parâmetros vetor média e matriz covariância associados à classe  $w_i$ .

$$\text{Vetor média } \mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{bmatrix}$$

$$\text{Matriz covariância } \Sigma_i = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \dots & \sigma_{i1p} \\ \sigma_{i21} & \sigma_{i22} & \dots & \sigma_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{ip1} & \sigma_{ip2} & \dots & \sigma_{ipp} \end{bmatrix}$$

onde  $\mu_{ij}$  é o valor esperado da característica  $j$  na classe  $i$ ;

$\sigma_{ijm}$  é a variância, ou a covariância, entre as características  $j$  e  $m$  da classe  $i$ .

Para estabelecer a regra de decisão do classificador de máxima verossimilhança gaussiana vamos considerar  $p(\mathbf{x} | w_i)$  uma distribuição normal  $p$ -variada. Por simplicidade iremos desconsiderar diferentes custos de erros de classificação ou, equivalentemente, estaremos atribuindo  $L(w_i, w_j) = 0$  para classificação correta e  $L(w_i, w_j) = 1$  para classificação errada.

$$P(w_i | \mathbf{x}) = P(w_i) \times p(\mathbf{x} | w_i)$$

$$P(w_i | \mathbf{x}) = P(w_i) \times \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)' \Sigma_i^{-1}(\mathbf{x} - \mu_i)\right\}$$

O termo  $(2\pi)^{p/2}$ , constante a todas classes  $w_i$ , pode ser desconsiderado na regra de classificação, ficando:

$$P(w_i | \mathbf{x}) = P(w_i) \times \frac{1}{|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)' \Sigma_i^{-1}(\mathbf{x} - \mu_i)\right\}$$

A forma final da função de discriminação baseada na máxima verossimilhança gaussiana fica mais simples se tomarmos o logaritmo natural da expressão anterior:

$$\ln[P(w_i | \mathbf{x})] = \ln[P(w_i)] - \frac{1}{2} \ln|\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

Por simplicidade de notação, consideraremos  $G_i(\mathbf{x}) = \ln[P(w_i | \mathbf{x})]$ .

Portanto, com a nova notação:

$$G_i(\mathbf{x}) = \ln[P(w_i)] - \frac{1}{2} \ln|\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

A regra de classificação para um dado padrão  $\mathbf{x}$ , será a seguinte:

$$\mathbf{x} \in w_i \quad \text{se} \quad G_i(\mathbf{x}) > G_j(\mathbf{x}) \quad \forall j \neq i$$

Se o pesquisador não dispuser de informações prévias sobre as probabilidades *a priori* de cada classe  $w_i$ , poderá utilizar o que chamamos de uma ‘*priori* não informativa’ considerando  $P(w_i) = 1/k$  para todas as classes  $w_1, w_2, \dots, w_k$ . Nesse caso, a função de discriminação fica mais simples e o fator comum  $1/2$  pode ser omitido.

$$G_i(\mathbf{x}) = -\ln|\Sigma_i| - (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

Podemos simplificar ainda mais a expressão, desde que haja razões suficientes para considerarmos a existência de uma matriz covariância comum a todas as classes  $w_i$ . Considerando  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$ , a função decisão toma a forma da distância de Mahalanobis.

$$G_i(\mathbf{x}) = -(\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i)$$

A estimação dos parâmetros será feita a partir de um conjunto de amostras de treinamento  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  devidamente rotuladas com as respectivas classes  $w_i$  as quais pertencem. Segundo McLachlan (1992, p. 54) e Johnson e Wichern (1982, p.146) os estimadores de máxima verossimilhança para os parâmetros desconhecidos  $\mu_i$  e  $\Sigma_i$  são, respectivamente, o vetor das médias amostrais  $\bar{\mathbf{X}}_i$  e a matriz covariância amostral  $\mathbf{S}_i$ , desde que corrigida pelo fator  $(n-1)/n$ . A estatística  $\mathbf{S}_i$ , apesar de não ser o estimador de máxima verossimilhança, é um estimador não-tendencioso de  $\Sigma_i$ .

### *3.2.1 Fronteiras de decisão no caso gaussiano*

A função de discriminação  $G_i(\mathbf{x})$ , para o caso onde são consideradas diferentes matrizes covariância  $\Sigma_i$  para cada classe, é conhecida como uma regra de discriminação quadrática. De acordo com Richards (1993, p.185), as fronteiras de decisão implementadas pelo classificador de máxima verossimilhança gaussiana são quadráticas e tomam a forma de parábolas, círculos ou elipses. As fronteiras de decisão entre as classes  $w_i$  e  $w_j$  são obtidas, simplesmente, igualando as funções de discriminação  $G_i(\mathbf{x}) = G_j(\mathbf{x})$ . Richards (1993, p.184) afirma que, tratando-se de classes espectrais, a separação entre pares de classes será feita através de superfícies no domínio multispectral. As classes espectrais são definidas por regiões no espaço multispectral onde os escores de suas funções discriminante são os maiores.

No caso homoscedástico, ou seja, sob a suposição de uma matriz covariância comum  $\Sigma$ , as superfícies de decisão entre pares de classes são lineares, descrevendo retas, planos ou hiperplanos.

### 3.3 Discriminação Logística

O modelo tradicional de regressão logística para variáveis-resposta binárias, bem como a extensão do modelo para três níveis, foram apresentados no capítulo 2. Agora será apresentada a generalização do modelo logístico para variáveis-resposta com  $k$  níveis ( $k > 2$ ), o que permitirá a sua utilização para discriminação entre  $k$  classes. Na discriminação logística a probabilidade de um dado padrão  $\mathbf{x}$  pertencer a uma das classes  $w_i$  é estimada diretamente por meio da seguinte expressão:

$$P(w_i | \mathbf{x}) = \frac{\exp(\beta_{i0} + \beta_i' \mathbf{x})}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j' \mathbf{x})} \quad \text{onde, } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \beta_i = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{ip} \end{bmatrix} \quad 1 \leq i \leq k-1$$

Considerando  $w_1, w_2, \dots, w_k$  classes exaustivas e exclusivas, podemos afirmar que  $\sum_{i=1}^k P(w_i | \mathbf{x}) = 1$ . Portanto, a probabilidade do padrão  $\mathbf{x}$  pertencer a  $k$ -ésima classe, denotada por  $P(w_k | \mathbf{x})$ , pode ser obtida por diferença:

$$P(w_k | \mathbf{x}) = 1 - \sum_{i=1}^{k-1} P(w_i | \mathbf{x})$$

O modelo logístico necessita da estimação de  $k-1$  vetores de parâmetros  $\beta_i$ , correspondentes a  $k-1$  classes presentes na imagem. A  $k$ -ésima classe é assumida como base. O logaritmo natural da razão entre as probabilidades do padrão  $\mathbf{x}$  pertencer às classes  $w_i$  ( $i=1, \dots, k-1$ ) e a probabilidade de pertencer ao nível base  $w_k$  são assumidas como sendo funções lineares. Portanto, teremos  $k-1$  funções *logit*  $g_i(\mathbf{x})$ :

$$g_i(\mathbf{x}) = \ln \left[ \frac{P(w_i | \mathbf{x})}{P(w_k | \mathbf{x})} \right] = \beta_{i0} + \beta_i' \mathbf{x} \quad 1 \leq i \leq k-1$$

McLachlan (1992, p.255) considera a suposição de linearidade nos parâmetros fundamental na abordagem logística e, por esse motivo, a chama de modelagem parcialmente paramétrica, pois apenas as funções *logit* estão sendo modeladas (consideradas lineares).

Para exemplificar a utilização da discriminação logística, vamos considerar um caso onde a variável resposta possui três níveis ( $k=3$ ), correspondentes a três diferentes classes  $w_1$ ,  $w_2$  e  $w_3$ . As probabilidades de um padrão  $\mathbf{x}$  pertencer a uma das classes  $w_i$  são calculadas da seguinte forma:

$$P(w_1 | \mathbf{x}) = \frac{\exp(\beta_{10} + \beta_1' \mathbf{x})}{1 + \sum_{i=1}^2 \exp(\beta_{i0} + \beta_i' \mathbf{x})}$$

$$P(w_2 | \mathbf{x}) = \frac{\exp(\beta_{20} + \beta_2' \mathbf{x})}{1 + \sum_{i=1}^2 \exp(\beta_{i0} + \beta_i' \mathbf{x})}$$

$$P(w_3 | \mathbf{x}) = \frac{1}{1 + \sum_{i=1}^2 \exp(\beta_{i0} + \beta_i' \mathbf{x})} = 1 - [P(w_1 | \mathbf{x}) + P(w_2 | \mathbf{x})]$$

Os vetores  $\beta_1$  e  $\beta_2$ , com dimensão  $p \times 1$  cada, e os interceptos  $\beta_{10}$  e  $\beta_{20}$  constituem os  $2(p+1)$  parâmetros do modelo a serem estimados. Vejamos as funções *logit* associadas a esse modelo:

$$g_1(\mathbf{x}) = \ln \left[ \frac{P(w_1 | \mathbf{x})}{P(w_3 | \mathbf{x})} \right] = \beta_{10} + \beta_1' \mathbf{x} = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1p}x_p$$

$$g_2(\mathbf{x}) = \ln \left[ \frac{P(w_2 | \mathbf{x})}{P(w_3 | \mathbf{x})} \right] = \beta_{20} + \beta_2' \mathbf{x} = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2p}x_p$$

$$g_3(\mathbf{x}) = \ln \left[ \frac{P(w_3 | \mathbf{x})}{P(w_3 | \mathbf{x})} \right] = 0$$

A utilização do modelo logístico para discriminação de classes pode ser direta. A regra de classificação para alocar um dado padrão  $\underline{x}$  numa das classes  $w_i$  é muito simples:

$$\mathbf{x} \in w_i \quad \text{se} \quad P(w_i | \mathbf{x}) > P(w_j | \mathbf{x}) \quad \forall j \neq i$$

Assim, como no caso do classificador gaussiano, a regra de classificação na discriminação logística pode ser vista num contexto bayesiano, com a inclusão de informações obtidas *a priori*. Além disso, é possível incluir dados referentes aos

custos de erros de classificação para se estabelecer uma regra de classificação baseada na minimização da função risco  $R(w_i | \mathbf{x})$ , conforme segue:

$$R(w_i | \mathbf{x}) = \sum_{j=1}^k L(w_i, w_j) \times P(w_j) \times P(w_j | \mathbf{x})$$

onde,

$P(w_i)$  = probabilidade *a priori* de um padrão pertencer à classe  $w_i$ .

$P(w_i | \mathbf{x})$  = probabilidade do padrão  $\mathbf{x}$  pertencer à classe  $w_i$ ,

$L(w_i, w_j)$  = custo de alocar o padrão  $\mathbf{x}$  na classe  $w_i$  quando a classe verdadeira é  $w_j$ .

$k$  = número de classes

Nesse caso, a regra de classificação aloca o padrão  $\mathbf{x}$  na classe  $w_i$  onde a função risco é mínima, da seguinte forma:

$$\mathbf{x} \in w_i \quad \text{se} \quad R(w_i | \mathbf{x}) < R(w_j | \mathbf{x}) \quad \forall j \neq i$$

### 3.3.1 Processo de estimação de parâmetros na discriminação logística

O processo de estimação dos parâmetros em regressão logística está baseado na maximização da função de verossimilhança  $\ell(\mathbf{x}, \beta)$ . Para tornar possível a realização desse procedimento são necessárias  $n$  amostras de treinamento  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , cujas classes a que pertencem são conhecidas.

Os vetores solução que maximizam a função  $\ell(\mathbf{x}, \beta)$  são aqueles que tornam máxima a probabilidade da particular amostra de treinamento  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  ter sido selecionada. Sendo assim, sob a hipótese da amostra ser representativa da população em estudo, estaremos obtendo um modelo que maximiza as chances de classificar todos padrões da população nas classes  $w_i$  as quais realmente pertencem.

Para apresentar a função de verossimilhança da discriminação logística, temos de criar  $k$  variáveis *dummy*, as quais chamaremos de  $y_1, y_2, \dots, y_k$  que assumem o valor 1 se o padrão pertence à classe correspondente, e zero em caso contrário, conforme mostra a tabela 1. A tabela 2 apresenta um exemplo de amostra de treinamento com a notação que será utilizada na função de verossimilhança, que inclui um índice  $h$  para individualizar as  $n$  observações amostrais.

TABELA 1

Variáveis *dummy*  $y_i$  para a construção da função de verossimilhança

<i>dummy</i>	$y_1$	$y_2$	...	$y_k$
Classe				
$w_1$	1	0	...	0
$w_2$	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$w_k$	0	0	...	1

TABELA 2

Exemplo de amostra de treinamento

Observação ( $h$ )	Classe a que pertence ( $y_{ih}$ )	Variáveis independentes			
		$x_{1h}$	$x_{2h}$	...	$x_{ph}$
1	$y_{i1}$	$x_{11}$	$x_{21}$	...	$x_{p1}$
2	$y_{i2}$	$x_{12}$	$x_{22}$	...	$x_{p2}$
3	$y_{i3}$	$x_{13}$	$x_{23}$	...	$x_{p3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$y_{in}$	$x_{1n}$	$x_{2n}$	...	$x_{pn}$

Utilizando as variáveis *dummy*, a função de verossimilhança  $\ell(\mathbf{x}, \beta)$  fica:

$$\ell(\mathbf{x}, \beta) = \prod_{h=1}^n \left( \frac{\exp(g_1(\mathbf{x}_h))}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{y_{1h}} \times \left( \frac{\exp(g_2(\mathbf{x}_h))}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{y_{2h}} \times \dots \times \left( \frac{\exp(g_k(\mathbf{x}_h))}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{y_{kh}}$$

Como  $\exp(g_k(\mathbf{x}_h))=1$ , pois  $g_k(\mathbf{x}_h)=0$  devido ao fato do nível  $k$  ser assumido como base, e  $\sum_{i=1}^k y_{ih} = 1$  para  $h$  fixo, podemos reescrever a função de

verossimilhança:

$$\ell(\mathbf{x}, \beta) = \prod_{h=1}^n \left( \frac{\exp(g_1(\mathbf{x}_h))}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{y_{1h}} \times \left( \frac{\exp(g_2(\mathbf{x}_h))}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{y_{2h}} \times \dots \times \left( \frac{1}{1 + \sum_{c=1}^{k-1} \exp(g_c(\mathbf{x}_h))} \right)^{1 - \sum_{j=1}^{k-1} y_{jh}}$$

onde  $g_i(\mathbf{x}_h) = \exp(\beta_{i0} + \beta_i' \mathbf{x}_h)$

É mais fácil trabalhar matematicamente com o logaritmo natural da função de verossimilhança, o que não interfere no processo de maximização. Chegamos, então, a uma expressão mais simples:

$$\ln \ell(\mathbf{x}, \beta) = \sum_{h=1}^n y_{1h} g_1(\mathbf{x}_h) + y_{2h} g_2(\mathbf{x}_h) + \dots + y_{(k-1)h} g_{k-1}(\mathbf{x}_h) - \ln \left( 1 + \sum_{c=1}^{k-1} g_c(\mathbf{x}_h) \right)$$

As equações de verossimilhança são encontradas tomando as primeiras derivadas parciais de  $\ln \ell(\mathbf{x}, \beta)$  em relação a cada um dos parâmetros  $\beta_{ij}$  do modelo.

Para maximizá-las igualamos cada equação a zero.

$$\frac{\partial \ell(\mathbf{x}, \beta)}{\partial \beta_{ij}} = 0 \quad 1 \leq i \leq k-1 \quad 0 \leq j \leq p$$

A forma geral das equações de verossimilhança é:

$$\frac{\partial \ell(\mathbf{x}, \beta)}{\partial \beta_{ij}} = \begin{cases} \sum_{h=1}^n \left( y_{1h} - \frac{e^{g_i(\mathbf{x}_h)}}{1 + \sum_{c=1}^{k-1} e^{g_c(\mathbf{x}_h)}} \right) = 0 & \text{se } 1 \leq i \leq k-1 \quad j = 0 \\ \sum_{h=1}^n \mathbf{x}_{jh} \left( y_{ih} - \frac{e^{g_i(\mathbf{x}_h)}}{1 + \sum_{c=1}^{k-1} e^{g_c(\mathbf{x}_h)}} \right) = 0 & \text{se } 1 \leq i \leq k-1 \quad 1 \leq j \leq p \end{cases}$$

onde  $g_i(\mathbf{x}_h) = \exp(\beta_{i0} + \beta_i' \mathbf{x}_h)$

A solução do sistema de equações de verossimilhança nos leva as estimativas para os  $k-1$  vetores de parâmetros  $\hat{\beta}_i$ . Tratando-se de equações claramente não lineares, necessitamos da utilização de métodos numéricos para encontrarmos uma solução. Esses processos são iterativos e estão disponíveis em alguns *softwares* estatísticos. No presente estudo utilizamos o procedimento CATMOD do sistema SAS que, conforme Allison (1999, p. 114), utiliza o método de Newton-Raphson, bastante rápido para convergência. Esse procedimento, descrito em SAS (1989, p. 405-517) e Spanos (1999, p.667), é próprio para modelagem de dados categóricos.

A técnica de regressão logística também pode ser realizada no SAS através dos procedimentos LOGISTIC e GENMOD. Entretanto, o procedimento CATMOD é o único capaz de obter as estimativas de máxima verossimilhança no caso de variável resposta multinomial.

O procedimento *Multinomial Logistic Regression* disponível no *software* SPSS versão 10.0 – descrito em Norusis e SPSS Inc. (1999, p. 65) – também pode ser utilizado com variáveis-resposta multinomiais. Apesar de apresentar uma interface bastante amigável ao usuário, os resultados obtidos com o SPSS em grandes amostras mostraram estimativas extremamente elevadas para os vetores de parâmetros  $\beta_i$ , o que nos levou a dar preferência ao SAS.

### 3.3.2 Testes de significância

O primeiro teste de significância importante em discriminação logística é o teste da razão de verossimilhança, onde testamos a hipótese de que pelo menos um dos parâmetros  $\beta_{ij}$  é diferente de zero (exceto os interceptos – parâmetros  $\beta_{i0}$ ). Esse teste faz uma comparação entre o valor da função de verossimilhança para o modelo contendo apenas os interceptos e a verossimilhança do modelo final com todos os parâmetros estimados. A estatística de teste  $D$ , chamada de *deviance*, tem uma distribuição qui-quadrado e é calculada da seguinte forma:

$$D = -2 \ln \left( \frac{\ell(\boldsymbol{\beta}_0)}{\ell(\mathbf{x}, \boldsymbol{\beta})} \right) = -2 \ln \ell(\boldsymbol{\beta}_0) - 2 \ln \ell(\mathbf{x}, \boldsymbol{\beta}) \sim \chi_{(k-1)p}^2$$

onde,

$\ell(\boldsymbol{\beta}_0)$  é o valor da função de verossimilhança apenas com os interceptos

$\ell(\mathbf{x}, \boldsymbol{\beta})$  é o valor da função de verossimilhança para o modelo final

$k$  é o número de classes

$p$  é o número de variáveis incluídas no modelo

Para a realização de testes de significância individuais para os parâmetros  $\beta_{i0}$  e  $\beta_{ij}$ , é bastante comum a utilização da bem-conhecida estatística de *Wald*, onde a hipótese nula é a de que o particular coeficiente  $\beta_{ij}$  é igual a zero. A estatística  $W$  de *Wald* é definida como o quadrado da razão entre a estimativa de máxima verossimilhança para o coeficiente e seu respectivo erro-padrão (EP). Essa estatística tem uma distribuição assintoticamente qui-quadrado com um único grau de liberdade:

$$W = \left( \frac{\hat{\beta}_{ij}}{EP(\hat{\beta}_{ij})} \right)^2 \sim \chi_1^2$$

Os erros-padrão das estimativas de máxima verossimilhança  $\hat{\beta}_{ij}$  são calculados a partir da matriz informação  $I(\beta)$ , obtida a partir do negativo do valor esperado da matriz de derivadas parciais de segunda ordem do logaritmo natural da função de verossimilhança  $\ell(\mathbf{x}, \beta)$ . A matriz  $I(\beta)$  é quadrada, com dimensão  $(k-1)(p+1)$ . Os termos que compõem a matriz são obtidos da seguinte forma:

$$I(\beta) = -E \left[ \frac{\partial^2 \ell(\mathbf{x}, \beta)}{\partial \beta_{ij} \partial \beta_{i'j'}} \right] \quad \text{se } \begin{matrix} 1 \leq i, i' \leq k-1 & 0 \leq j, j' \leq p \end{matrix}$$

A matriz covariância assintótica  $\hat{\Sigma}(\beta)$  para as estimativas de máxima verossimilhança é obtida tomando-se o inverso da matriz informação  $I(\beta)$ :

$$\hat{\Sigma}(\beta) = I(\beta)^{-1}$$

Apesar do teste de Wald ser muito conhecido, Hauck e Donner (1977, p. 853) não aconselham sua utilização porque observaram no mesmo um comportamento aberrante, apresentando freqüentes falhas na rejeição de coeficientes que deveriam ser significativos. Os autores sugerem a utilização do teste da razão de verossimilhança para verificar a significância de cada variável  $x_i$  individualmente. A estatística de teste  $G$  faz uma comparação entre o valor da função de verossimilhança com e sem a variável  $x_i$ :

$$G = -2 \ln \left[ \frac{\ell(\mathbf{x}, \beta)_{/x_i}}{\ell(\mathbf{x}, \beta)} \right] \sim \chi_1^2$$

onde,

$\ell(\mathbf{x}, \beta)_{/x_i}$  é o valor da função de verossimilhança sem a variável  $x_i$

$\ell(\mathbf{x}, \beta)$  é o valor da função de verossimilhança com a variável  $x_i$

### 3.3.3 Interpretação dos parâmetros

A interpretação dos parâmetros estimados no modelo de regressão logística torna-se similar ao caso da regressão múltipla tradicional. Como vimos anteriormente, no caso de uma variável resposta com  $k$  níveis, o  $k$ -ésimo nível será assumido como base e, portanto, podemos estabelecer  $k-1$  funções *logit*, contrastando cada nível contra o nível base, conforme segue:

$$g_i(\mathbf{x}) = \ln \left[ \frac{P(w_i | \mathbf{x})}{P(w_k | \mathbf{x})} \right] = \beta_{i0} + \beta_i' \mathbf{x} \quad 1 \leq i \leq k-1$$

Aplicando a função exponencial nos dois lados da igualdade, temos:

$$e^{g_i(\mathbf{x})} = \frac{P(w_i | \mathbf{x})}{P(w_k | \mathbf{x})} = e^{\beta_{i0} + \beta_i' \mathbf{x}} \quad 1 \leq i \leq k-1$$

Assim, temos que um aumento de uma unidade na variável  $x_j$  causará um aumento de  $e^{\beta_{ij}}$  unidades na razão entre as probabilidades do padrão  $\mathbf{x}$  pertencer a

classe  $w_i$  em relação à classe  $w_k$ . Portanto, quando  $x_j$  aumenta em uma unidade, a classe  $w_i$  torna-se  $e^{\beta_{ij}}$  vezes mais provável que  $w_k$ .

### 3.4 Material Utilizado

Os recursos computacionais (*hardware*) necessários para a realização do trabalho foram um computador AMD K6-2 com 64Mb de memória RAM e impressora jato de tinta Epson 480SX com resolução de 720dpi. Os principais *softwares* foram o programa matemático MATLAB versão 5.3, o pacote estatístico SAS versão 6.0 e o programa de processamento de imagens Multispec versão 1.2. O *software* estatístico SPSS versão 10.0 foi utilizado para elaboração de gráficos tridimensionais.

As duas imagens digitais utilizadas nesse estudo foram coletadas pelos sensores TM e AVIRIS. A imagem do satélite Landsat 5, sensor TM, coletada em 1992 cobre uma extensa área do litoral gaúcho, caracterizada pela presença de muitas lagoas e da serra geral. Já a imagem coletada pelo sensor aerotransportado AVIRIS, obtida em junho de 1992, cobre uma área agrícola do estado norte-americano de Indiana.

## 4. RESULTADOS E DISCUSSÕES

### 4.1 Comparação entre os Processos de Estimação

Ambos os métodos – discriminação logística e máxima verossimilhança gaussiana – são supervisionados e paramétricos, pois necessitam da estimação de parâmetros a partir de um conjunto de amostras de treinamento.

No método da máxima verossimilhança gaussiana cada uma das  $k$  classes necessita da estimação de  $p$  parâmetros para o vetor média e  $\frac{p(p+1)}{2}$  parâmetros para a matriz covariância. Portanto, considerando a discriminação entre  $k$  classes pelo método da máxima verossimilhança, com diferentes matrizes covariância para cada classe, necessitamos da estimação de  $k\left(p + \frac{p(p+1)}{2}\right)$  parâmetros. Sob a suposição de igualdade das  $k$  matrizes covariância, a função de discriminação toma a forma da distância de Mahalanobis, e o número de parâmetros necessários decai muito, para  $kp + \frac{p(p+1)}{2}$ .

O modelo utilizado para discriminação logística é freqüentemente rotulado como um modelo robusto. Um dos motivos desse rótulo deve-se o fato do modelo necessitar da estimação de um número razoavelmente pequeno de parâmetros, o que permite, ao menos teoricamente, uma redução no número de amostras de treinamento. Em discriminação logística necessita-se de apenas  $p+1$  parâmetros para cada uma das  $k-1$  classes de padrões.

A figura 1 apresenta a relação entre o número de características ( $p$ ) e o número de parâmetros necessários em cada um dos classificadores – máxima verossimilhança gaussiana com diferentes matrizes covariância, distância de Mahalanobis e discriminação logística – considerando  $k=3$  classes.

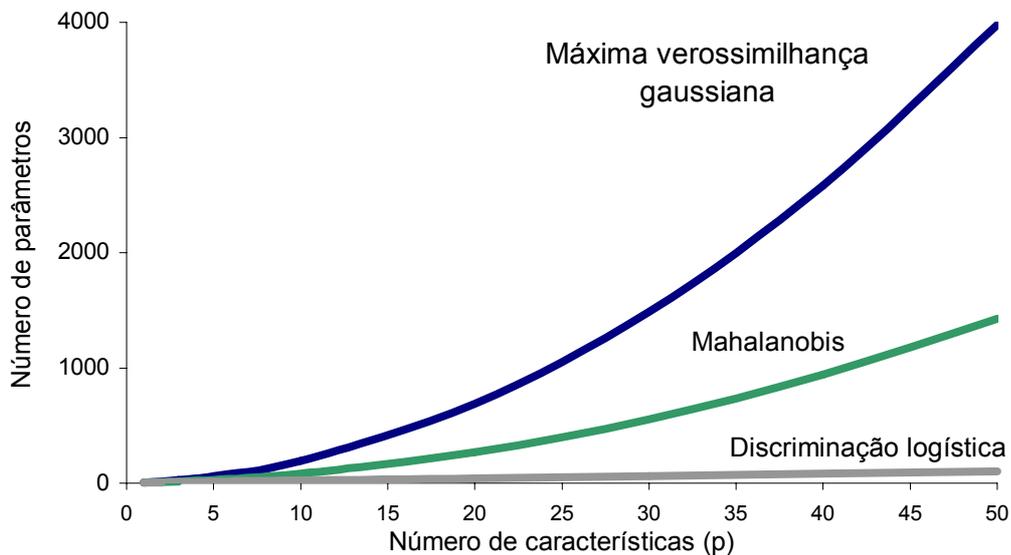


FIGURA 2: Número de parâmetros a ser estimado para discriminação de três classes ( $k=3$ ) em função do número de características ( $p$ )

Enquanto o número de parâmetros cresce linearmente em função do número de características na discriminação logística, o crescimento é quadrático nos classificadores baseados no modelo gaussiano. A taxa de crescimento é mais alta no caso da máxima verossimilhança gaussiana com diferentes matrizes covariância.

Apesar da vantagem em relação ao número de parâmetros, o processo de estimação na discriminação logística é mais complexo, pois necessita da utilização de métodos numéricos para obtenção das estimativas de máxima verossimilhança. Os métodos numéricos são procedimentos iterativos que só tornam-se viáveis com a utilização de computadores. A presença de colinearidade – correlação entre as características  $x_j$  – geralmente ocasiona superestimativas para os parâmetros do modelo e/ou estimativas extremamente elevadas para os erros-padrão.

Os classificadores da máxima verossimilhança gaussiana e a distância de Mahalanobis necessitam de boas estimativas dos vetores média e da(s) matriz(es) covariância, o que só é possível com quantidades razoáveis de amostras de treinamento. O aumento na resolução espectral dos atuais sistemas de sensoriamento remoto tem ocasionado um considerável aumento no número de bandas espectrais e, conseqüentemente, o tamanho amostral necessário para treinamento de classificadores supervisionados deve ser suficientemente grande. Assim, quando trabalhamos com classes espectrais presentes em imagens digitais, o número de parâmetros pode facilmente tornar-se muito alto, a ponto de inviabilizar o processo de classificação.

## 4.2 Resultados Obtidos por Simulação

Com o objetivo de estudar o comportamento dos classificadores frente a variáveis com diferentes distribuições probabilísticas, foi realizado um estudo de simulação a partir de conjuntos de dados gerados no *software* MATLAB. Esse tipo de estudo é muito importante para um melhor entendimento do funcionamento dos classificadores, visto que a classificação é realizada sob condições totalmente controladas.

### 4.2.1 Classes uniformemente distribuídas

Utilizando o gerador de números aleatórios do MATLAB foram simulados 30 mil vetores  $\mathbf{x}' = [x_1 \quad x_2]$  divididos em três classes:  $w_1$ ,  $w_2$  e  $w_3$ . Todas classes apresentam mesma variabilidade interna, mas possuem diferentes vetores média, conforme podemos verificar na figura 3 e na tabela 3.

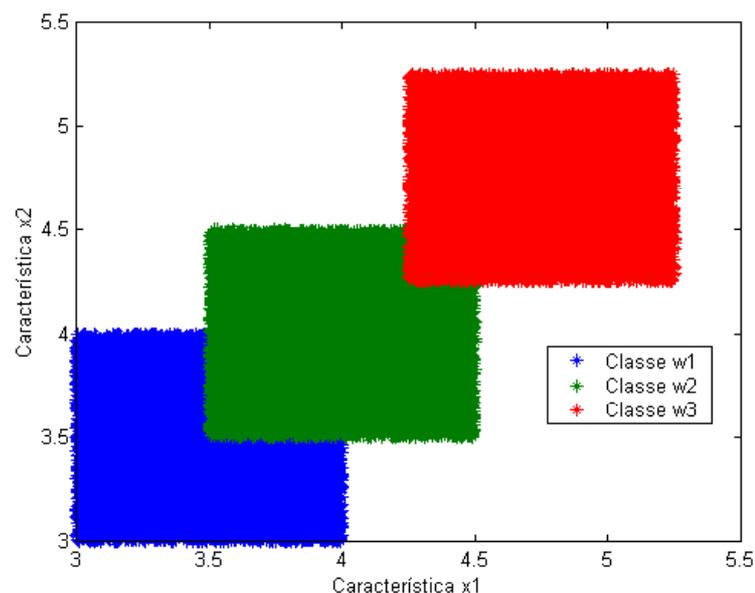


FIGURA 3: Comportamento das classes  $w_1$ ,  $w_2$  e  $w_3$  com relação às características  $x_1$  e  $x_2$ .

TABELA 3  
Vetores média e matrizes covariância por classe

<i>Classe</i>	Vetor média	Matriz covariância
$w_1$	$\mu_1 = \begin{bmatrix} 3,50 \\ 3,50 \end{bmatrix}$	$\Sigma_1 = \begin{bmatrix} 1/12 & 0 \\ 0 & 1/12 \end{bmatrix}$
$w_2$	$\mu_2 = \begin{bmatrix} 4,00 \\ 4,00 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 1/12 & 0 \\ 0 & 1/12 \end{bmatrix}$
$w_3$	$\mu_3 = \begin{bmatrix} 4,75 \\ 4,75 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 1/12 & 0 \\ 0 & 1/12 \end{bmatrix}$

Após extrair uma amostra aleatória de 300 vetores para cada classe, totalizando 900 observações, procedemos com a estimação dos parâmetros necessários na discriminação logística e na máxima verossimilhança gaussiana. Os parâmetros do modelo logístico foram estimados no *software* SAS, sendo que a classificação foi posteriormente realizada no MATLAB. Devido a facilidade de programação, a classificação por máxima verossimilhança gaussiana foi diretamente realizada no MATLAB.

Os parâmetros estimados no *software* SAS permitem escrever equações que estimam diretamente a probabilidade  $P(w_i | \mathbf{x})$ . Essas equações são:

$$P(w_1 | \mathbf{x}) = \frac{\exp\{143,5 - 16,8072x_1 - 17,6038x_2\}}{1 + \exp\{143,5 - 16,8072x_1 - 17,6038x_2\} + \exp\{102,1 - 11,2907x_1 - 12,0303x_2\}}$$

$$P(w_2 | \mathbf{x}) = \frac{\exp\{102,1 - 11,2907x_1 - 12,0303x_2\}}{1 + \exp\{143,5 - 16,8072x_1 - 17,6038x_2\} + \exp\{102,1 - 11,2907x_1 - 12,0303x_2\}}$$

$$P(w_3 | \mathbf{x}) = 1 - P(w_1 | \mathbf{x}) - P(w_2 | \mathbf{x})$$

Para exemplificar o funcionamento das equações, admita três padrões –  $\mathbf{x}_1' = [3,25 \ 3,25]$ ;  $\mathbf{x}_2' = [4,25 \ 4,25]$  e  $\mathbf{x}_3' = [4,75 \ 4,75]$  – que claramente pertencem às classes  $w_1$ ,  $w_2$  e  $w_3$ . As probabilidades condicionais estimadas são:

$$\begin{aligned}
 P(w_1 | \mathbf{x}_1) &= 0,99531 & P(w_1 | \mathbf{x}_2) &= 0,00307 & P(w_1 | \mathbf{x}_3) &= 2,16 \times 10^{-9} \\
 P(w_2 | \mathbf{x}_1) &= 0,00469 & P(w_2 | \mathbf{x}_2) &= 0,94900 & P(w_2 | \mathbf{x}_3) &= 0,00017 \\
 P(w_3 | \mathbf{x}_1) &= 1,76 \times 10^{-14} & P(w_3 | \mathbf{x}_2) &= 0,04793 & P(w_3 | \mathbf{x}_3) &= 0,99983
 \end{aligned}$$

O modelo discriminou satisfatoriamente as três classes de padrões, considerando esses valores de  $\mathbf{x}$ . A figura 4 mostra a saída do procedimento CATMOD do *software* SAS, contendo as estimativas e testes para os parâmetros.

The SAS System						
Parameter Estimates						
Iteration	1	2	3	4	5	6
0	0	0	0	0	0	0
1	16.3628	10.3566	-1.9321	-1.4087	-2.0771	-1.0981
2	32.2196	18.7213	-3.8807	-2.2230	-4.0182	-2.0756
3	55.2970	31.0601	-6.7148	-3.5888	-6.8683	-3.4984
4	80.7619	46.2582	-9.8392	-5.3007	-9.9591	-5.2487
5	104.0174	63.8233	-12.6199	-7.2787	-12.7130	-7.2841
6	123.5884	82.2259	-14.7987	-9.2869	-15.0568	-9.4866
7	137.6721	96.2786	-16.2493	-10.7327	-16.8326	-11.2590
8	142.9516	101.5589	-16.7595	-11.2430	-17.5320	-11.9585
9	143.4686	102.0759	-16.8068	-11.2904	-17.6032	-12.0297
10	143.4729	102.0803	-16.8072	-11.2907	-17.6038	-12.0303

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE			
Source	DF	Chi-Square	Prob
INTERCEPT	2	195.29	0.0000
X1	2	117.41	0.0000
X2	2	129.90	0.0000

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES					
Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	143.5	14.9226	92.44	0.0000
	2	102.1	14.5179	49.44	0.0000
X1	3	-16.8072	2.0690	65.99	0.0000
	4	-11.2907	1.9800	32.52	0.0000
X2	5	-17.6038	2.3171	57.72	0.0000
	6	-12.0303	2.2495	28.60	0.0000

FIGURA 4: Saída do procedimento CATMOD do *software* estatístico SAS

Ainda explorando o modelo logístico, as figuras 5 a 7 apresentam o comportamento das probabilidades  $P(w_i | \mathbf{x})$  em função das variações de  $x_1$  e  $x_2$ . Percebe-se claramente as regiões de maior probabilidade para cada classe de padrões.

As figuras 5 e 7 revelam o característico formato em 'S' da curva logística, enquanto a figura 6 mostra uma superfície bastante pronunciada na região referente à classe  $w_2$ .

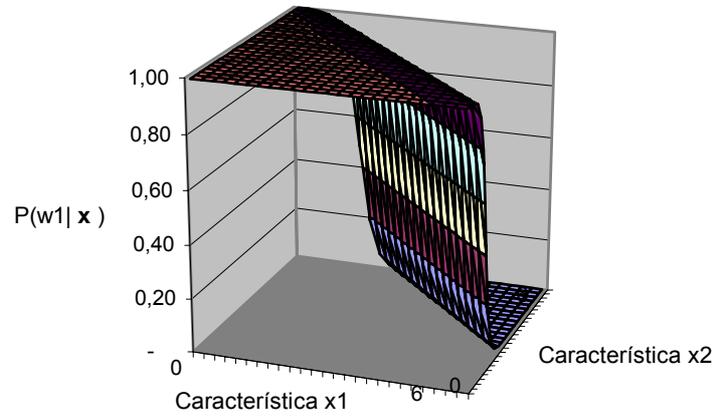


FIGURA 5: Probabilidade condicional  $P(w_1 | \mathbf{x})$  em função de  $x_1$  e  $x_2$ .

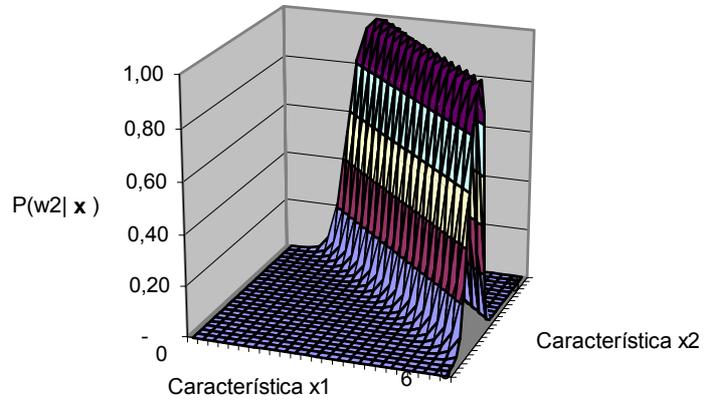


FIGURA 6: Probabilidade condicional  $P(w_2 | \mathbf{x})$  em função de  $x_1$  e  $x_2$ .

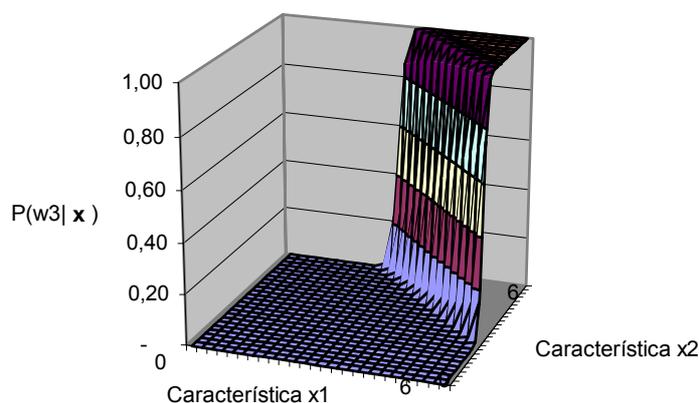


FIGURA 7: Probabilidade condicional  $P(w_3 | \mathbf{x})$  em função de  $x_1$  e  $x_2$ .

Para a realização da classificação por máxima verossimilhança gaussiana foram necessárias estimativas dos vetores média e das matrizes covariância de cada classe. A tabela 4 apresenta as estimativas baseadas na amostra de 900 padrões.

TABELA 4

Estimativas para os vetores média e matrizes covariância por classe

<i>Classe</i>	Vetor média	Matriz covariância
$w_1$	$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3,5212 \\ 3,5074 \end{bmatrix}$	$\hat{\Sigma}_1 = \begin{bmatrix} 0,088 & -0,001 \\ -0,001 & 0,085 \end{bmatrix}$
$w_2$	$\bar{\mathbf{x}}_2 = \begin{bmatrix} 3,9725 \\ 3,9986 \end{bmatrix}$	$\hat{\Sigma}_2 = \begin{bmatrix} 0,085 & 0,010 \\ 0,010 & 0,087 \end{bmatrix}$
$w_3$	$\bar{\mathbf{x}}_3 = \begin{bmatrix} 4,7490 \\ 4,7592 \end{bmatrix}$	$\hat{\Sigma}_3 = \begin{bmatrix} 0,087 & 0,009 \\ 0,009 & 0,081 \end{bmatrix}$

Como as matrizes covariância são praticamente iguais nas três diferentes classes, as funções de discriminação  $G_i(\mathbf{x})$  tomaram a forma da distância de Mahalanobis:

$$G_i(\mathbf{x}) = -(\mathbf{x} - \bar{\mathbf{x}}_i)' \hat{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \quad i = 1, 2 \text{ e } 3$$

As matrizes covariância sugerem independência entre as características  $x_1$  e  $x_2$ , pois os elementos fora da diagonal principal são praticamente nulos. Admitindo ausência de covariância, o classificador gaussiano tornou-se equivalente à distância euclidiana. A figura 8 apresenta as fronteiras de decisão para discriminação das três classes que, como era de se esperar, são lineares.

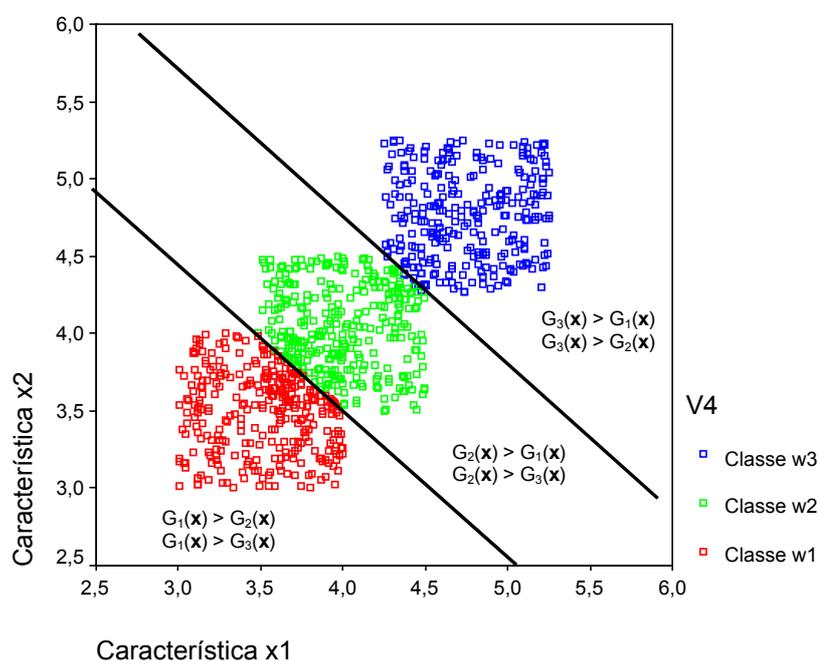


FIGURA 8: Fronteiras de decisão do classificador de máxima verossimilhança gaussiana

Observando a figura 3 percebe-se que a taxa esperada de erro de classificação pode ser obtida simplesmente computando as áreas de intersecção entre as classes e multiplicando-as por 1/2. A tabela 5 apresenta as taxas esperadas de erros de classificação e a tabela 6 resume os resultados obtidos na classificação dos 30 mil padrões que constituem a população.

TABELA 5

Taxa esperada de erros de classificação para as classes uniformemente distribuídas

Classe	Área de intersecção	Taxa esperada de erro de classificação	Taxa esperada de classificação correta
$w_1$	1/4	12,5%	87,5%
$w_2$	5/16	15,6%	84,4%
$w_3$	1/16	03,1%	96,9%
<i>Geral</i>	5/48	10,4%	89,6%

TABELA 6

Tabela de classificação por método

Classe verdadeira	Classificação por discriminação logística			Classificação por máxima verossimilhança gaussiana		
	$w_1$	$w_2$	$w_3$	$w_1$	$w_2$	$w_3$
$w_1$	85,9%	14,1%	0,0%	87,2%	12,8%	0,0%
$w_2$	10,7%	86,4%	2,9%	11,9%	84,8%	3,3%
$w_3$	0,0%	3,3%	96,7%	0,0%	2,9%	97,1%
	Acurácia geral =89,7%			Acurácia geral = 89,7%		

O desempenho de ambos classificadores foi praticamente idêntico, atingindo taxas gerais de classificação correta de 89,7%. O resultado indica que, para classes uniformemente distribuídas, com homogeneidade de variância e covariâncias nulas, o desempenho dos classificadores tende a ser similar. De fato, as fronteiras de decisão de ambos classificadores ficaram espacialmente muito próximas e foram definidas por retas.

#### 4.2.2 Classes normalmente distribuídas

Conforme descrito no capítulo 3, o classificador gaussiano assume que as distribuições de probabilidade  $P(\mathbf{x}|w_i)$  seguem uma distribuição normal  $p$ -variada, onde  $p$  denota o número de características consideradas no problema de classificação. Passamos, então, a simular dados com distribuição normal multivariada tendo como objetivo verificar o comportamento da discriminação logística frente a dados normalmente distribuídos.

Utilizando novamente o gerador de números aleatórios do MATLAB foram simulados 30 mil vetores  $\mathbf{x}' = [x_1 \ x_2]$ , divididos em três classes:  $w_1$ ,  $w_2$  e  $w_3$ . Cada classe possui diferentes vetores média e matrizes covariância, com exceção das classes  $w_2$  e  $w_3$  que propositalmente compartilham do mesmo vetor média. A tabela 7 apresenta os parâmetros utilizados na simulação e as figuras 9 e 10 ilustram graficamente os dados gerados.

TABELA 7

Descrição das classes  $w_1$ ,  $w_2$  e  $w_3$

Classe	Vetor média	Matriz covariância
$w_1$	$\mu_1 = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$	$\Sigma_1 = \begin{bmatrix} 0,6 & 0,3 \\ 0,3 & 0,4 \end{bmatrix}$
$w_2$	$\mu_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 0,8 & 0,6 \\ 0,6 & 0,7 \end{bmatrix}$
$w_3$	$\mu_3 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 0,8 & -0,7 \\ -0,7 & 0,7 \end{bmatrix}$

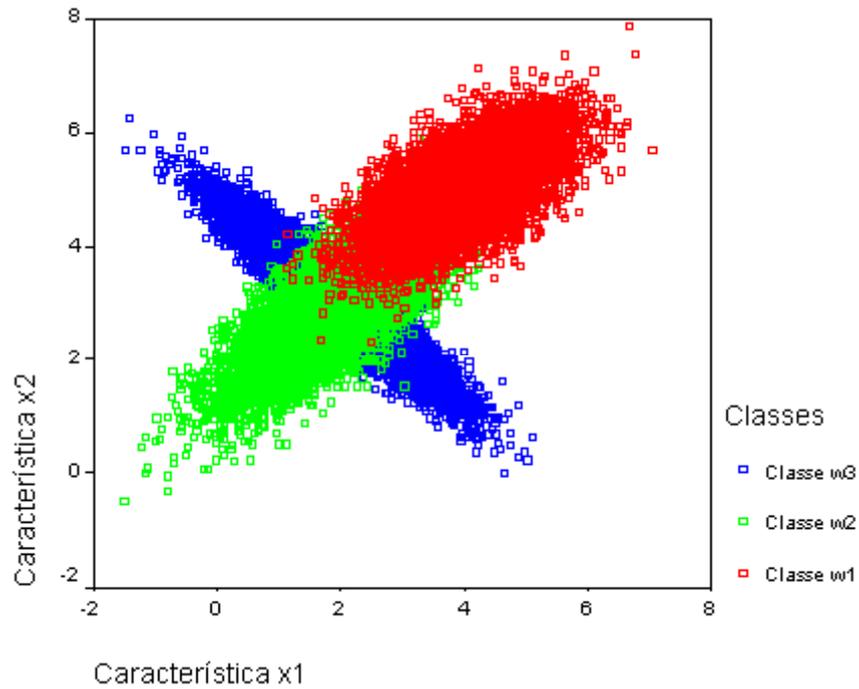


FIGURA 9: Comportamento das classes  $w_1$ ,  $w_2$  e  $w_3$  com relação às características  $x_1$  e  $x_2$ .

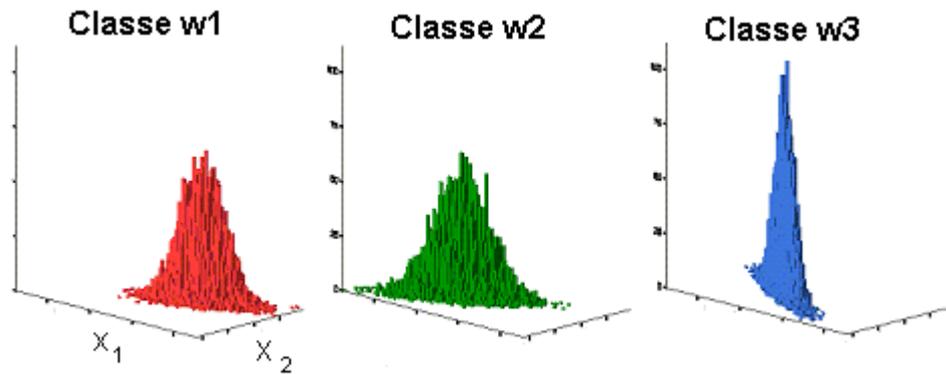


FIGURA 10: Histogramas bidimensionais ( $x_1$  e  $x_2$ ) para as classes  $w_1$ ,  $w_2$  e  $w_3$

O processo de estimação de parâmetros foi realizado a partir de um conjunto de 300 amostras de treinamento – 100 para cada classe – escolhidas de maneira aleatória. Os resultados do processo de estimação de parâmetros, realizado no *software* SAS, são apresentados na figura 11.

The SAS System

CATMOD PROCEDURE

Parameter Estimates						
Iteration	1	2	3	4	5	6
0	0	0	0	0	0	0
1	-4.1359	0.9622	0.7532	0.1166	0.5771	-0.3190
2	-7.3536	1.0356	1.0283	-0.0122	1.0968	-0.3027
3	-10.1636	1.2616	1.2383	-0.0493	1.5864	-0.3543
4	-12.2583	1.3157	1.3119	-0.0600	2.0115	-0.3665
5	-13.1171	1.3240	1.3200	-0.0611	2.2017	-0.3687
6	-13.2208	1.3246	1.3200	-0.0611	2.2253	-0.3689
7	-13.2221	1.3246	1.3200	-0.0611	2.2256	-0.3689

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	2	49.03	0.0000
X1	2	13.81	0.0010
X2	2	24.33	0.0000

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-13.2221	2.0593	41.22	0.0000
	2	<b>1.3246</b>	<b>0.5803</b>	<b>5.21</b>	<b>0.0225</b>
X1	3	1.3200	0.3694	12.77	0.0004
	4	<b>-0.0611</b>	<b>0.1695</b>	<b>0.13</b>	<b>0.7183</b>
X2	5	2.2256	0.5362	17.23	0.0000
	6	<b>-0.3689</b>	<b>0.1659</b>	<b>4.95</b>	<b>0.0262</b>

FIGURA 11: Saída do procedimento CATMOD do *software* estatístico SAS

A discriminação logística necessita da estimação de  $k-1$  vetores de parâmetros, correspondentes as  $k-1$  funções *logit*. Observando a saída do *software* SAS, percebe-se que os parâmetros da segunda função *logit* (destacados em negrito) não foram tão significativos quanto os parâmetros da primeira função. Isso ocorreu devido à incapacidade do modelo logístico em separar as classes  $w_2$  e  $w_3$ . Como as fronteiras de decisão da discriminação logística são lineares, a separação entre as classes que compartilham do mesmo vetor média tornou-se bastante deficiente, conforme ilustra a figura 12.

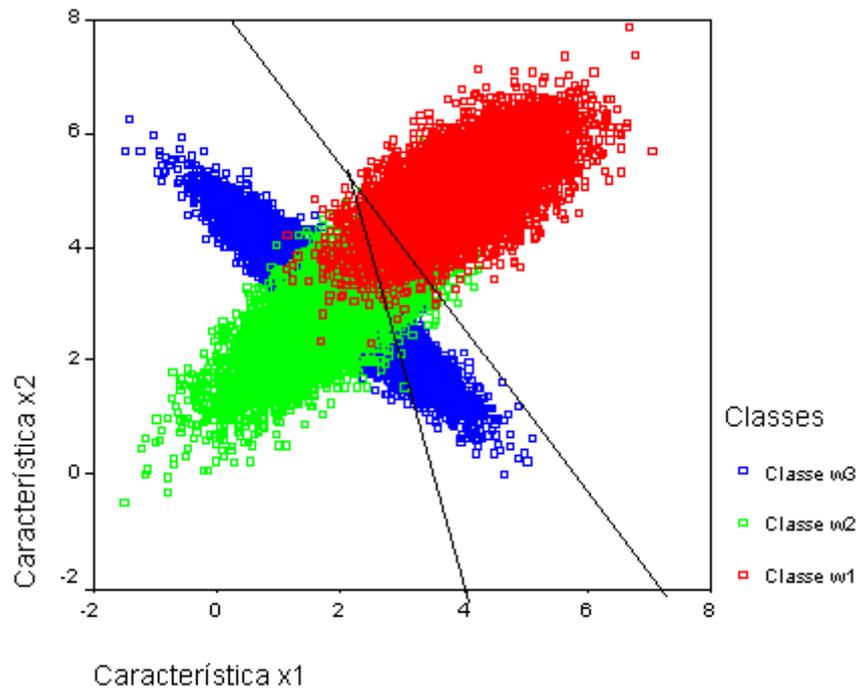


FIGURA 12: Fronteiras de decisão para o modelo de discriminação logística

O classificador de máxima verossimilhança gaussiana, por sua vez, gerou fronteiras de decisão quadráticas que possibilitaram maior separabilidade entre as classes. As formas elípticas das fronteiras de decisão do classificador gaussiano permitiram chegar a resultados muito superiores aos da discriminação logística, conforme podemos perceber na figura 13 e na tabela 8.

O resultado obtido não é inesperado, pois em situações de normalidade multivariada o classificador gaussiano é ótimo. Nesse caso específico, todas as funções densidade  $P(\mathbf{x}|w_i)$  são seguramente normais bivariadas. Como foram consideradas diferentes matrizes covariância para cada classe  $w_i$ , fez-se necessário estimar um total 15 parâmetros.

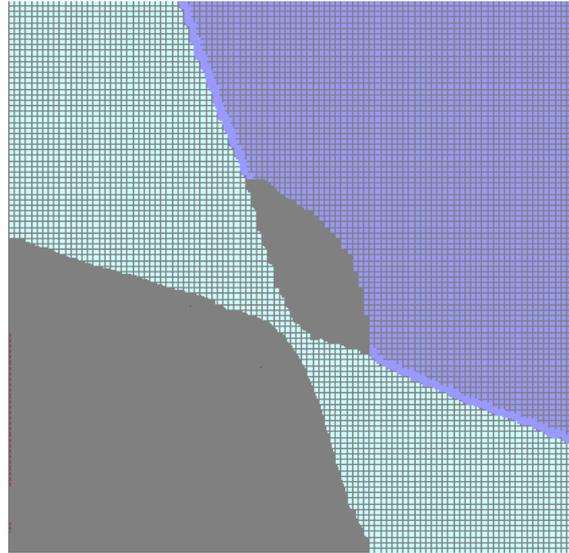


FIGURA 13: Ilustração das fronteiras de decisão no caso da máxima verossimilhança gaussiana

TABELA 8

Tabela de classificação por método baseada em 30.000 padrões

Classe verdadeira	Classificação por discriminação logística			Classificação por máxima verossimilhança gaussiana		
	$w_1$	$w_2$	$w_3$	$w_1$	$w_2$	$w_3$
$w_1$	<b>92,9%</b>	0,2%	7,0%	<b>94,1%</b>	5,6%	0,3%
$w_2$	9,2%	<b>60,8%</b>	30,0%	10,2%	<b>67,4%</b>	22,5%
$w_3$	0,0%	64,2%	<b>35,8%</b>	0,0%	6,2%	<b>93,8%</b>
	Acurácia geral = <b>63,2%</b>			Acurácia geral = <b>85,1%</b>		

#### 4.3 Resultados Obtidos com Imagens Digitais

Após observar o comportamento dos classificadores frente a situações controladas, o desempenho dos mesmos foi comparado em condições mais próximas da realidade. Dois segmentos de cenas coletadas pelos sensores remotos TM e AVIRIS foram utilizados.

Foram trabalhados apenas atributos espectrais. Portanto, os dados de entrada para os classificadores foram os valores dos contadores digitais registrados para os  $n$  pixels das imagens nas  $p$  bandas espectrais disponibilizadas pelos sensores.

#### *4.3.1 Classificação de imagem Landsat-TM*

Iniciamos com a classificação de um segmento de uma cena obtida pelo satélite Landsat 5, sensor TM, composta de 436 linhas, 535 colunas e 6 bandas espectrais. O número total de pixels da imagem é superior a 230 mil. Podemos visualizar na figura 14 que a imagem é basicamente composta de três classes bem diferenciadas: água, culturas e vegetação.



FIGURA 14: Segmento de uma cena Landsat-TM, composição colorida 5-4-2 (R-G-B)

As três classes apresentam comportamentos espectrais visualmente diferentes, em especial nas bandas espectrais 4 e 5, conforme podemos verificar na figura 15. Por esse motivo, espera-se que os classificadores apresentem baixos percentuais de erros de classificação

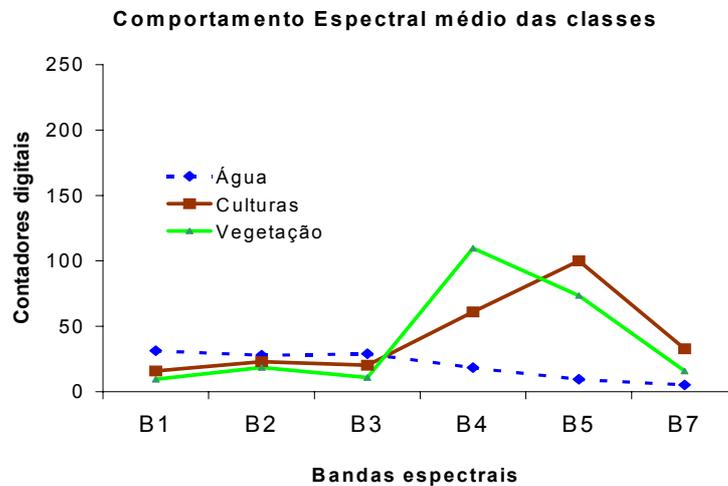


FIGURA 15: Comportamento espectral médio das classes nas seis bandas espectrais<sup>2</sup>

Ao contrário das situações anteriores, onde trabalhamos com apenas duas características, os pixels agora serão vistos como vetores num espaço a seis dimensões. Sendo assim, não há como construirmos gráficos a seis dimensões para avaliarmos o comportamento das classes. Os histogramas unidimensionais podem ajudar nessa tarefa, conforme podemos observar na figura 16 que apresenta os histogramas para as três classes na banda TM3 .

O processo de estimação dos parâmetros, tanto para a discriminação logística como para o método gaussiano, foi realizado sobre uma amostra de pouco mais de mil pixels devidamente rotulados. O modelo logístico necessitou da estimação de 14 parâmetros  $\beta_{ij}$ , enquanto que o classificador da máxima verossimilhança gaussiana precisou de 18 parâmetros para os vetores média e mais 63 para as matrizes covariância, totalizando 81 parâmetros.

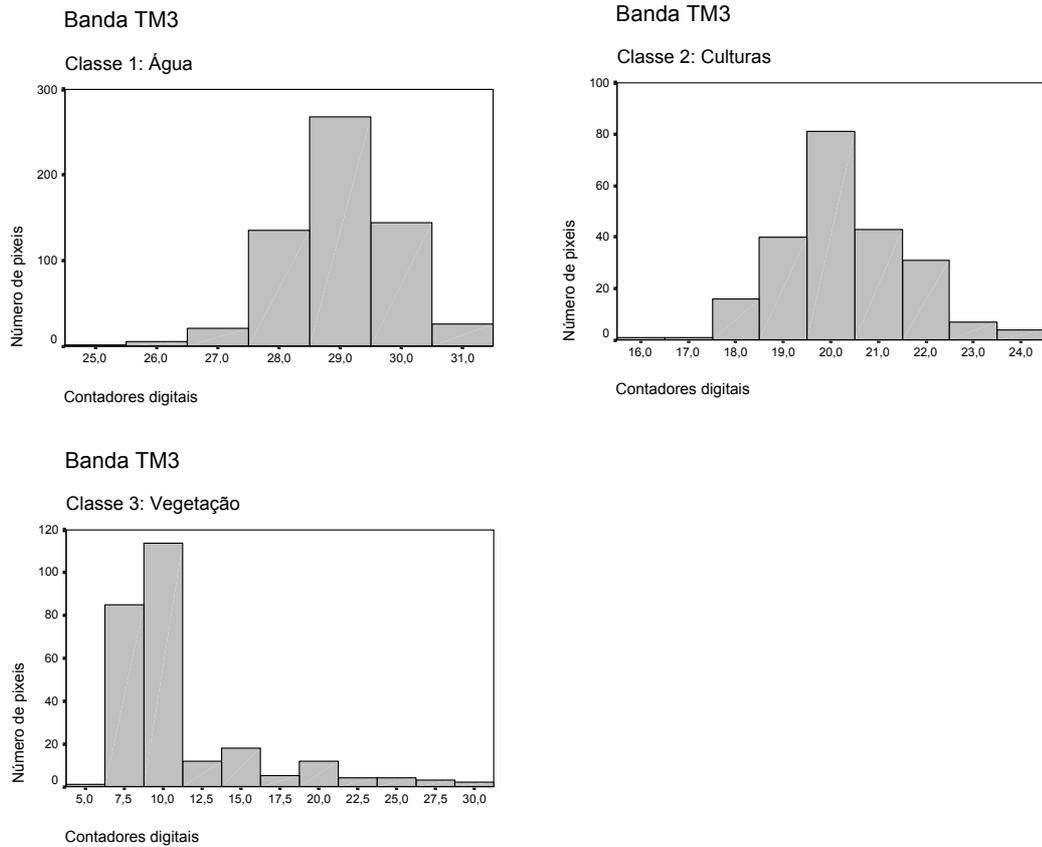


FIGURA 16: Histogramas da banda TM-3 nas três bandas espectrais

As equações que estimam as probabilidades  $P(w_i | \mathbf{x})$  na discriminação logística foram seguintes:

$$P(w_i | \mathbf{x}) = \frac{e^{g_i(\mathbf{x})}}{1 + \sum_{c=1}^2 e^{g_c(\mathbf{x})}}$$

Classes:  $w_1 = \text{Água}$ ;  $w_2 = \text{Culturas}$ ;  $w_3 = \text{Vegetação}$

$$g_1(x) = 113.6 - 0.1581 * x_1 + 1.1984 * x_2 - 0.2760 * x_3 - 1.7899 * x_4 - 0.1227 * x_5 + 0.1929 * x_7$$

$$g_2(x) = 141.3 - 0.7992 * x_1 + 2.3858 * x_2 - 1.9673 * x_3 - 3.1077 * x_4 + 0.9979 * x_5 + 0.1598 * x_7$$

<sup>2</sup> A banda espectral TM6, ou banda termal, não é utilizada por possuir diferente resolução espacial.

Antes de proceder a classificação pelo método da máxima verossimilhança gaussiana, a suposição de normalidade foi verificada. Mesmo que os histogramas sugiram dados normalmente distribuídos, testes de aderência de *Kolmogorov-Smirnov*, realizados para as seis bandas espectrais nas três classes, indicaram o contrário. Esses resultados univariados levam a crer que a hipótese de normalidade multivariada não seja plausível, indicando que a situação não é muito favorável ao classificador gaussiano. As figuras 17 e 18 mostram as imagens temáticas geradas a partir das classificações e a tabela 9 resume o resultado da classificação, baseado em aproximadamente duas mil amostras de teste.

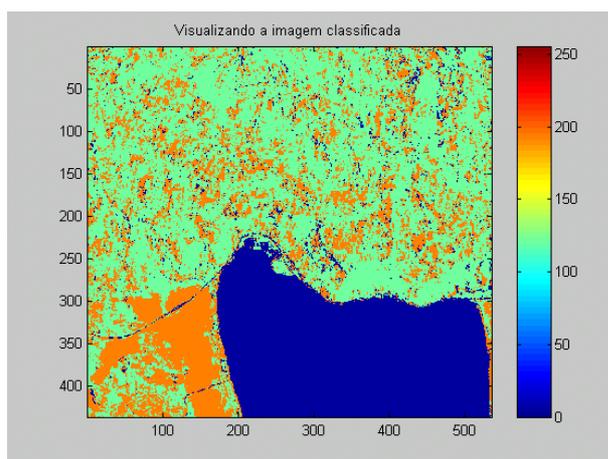


FIGURA 17: Imagem temática gerada a partir da classificação por discriminação logística

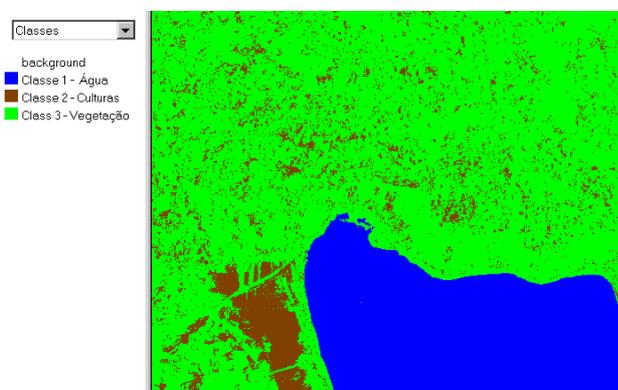


FIGURA 18: Imagem temática gerada a partir da classificação por máxima verossimilhança gaussiana

TABELA 9

Tabela de classificação baseada em 1.992 amostras de teste

Classe Verdadeira	Classificação por discriminação logística			Classificação por máxima verossimilhança gaussiana		
	Água	Culturas	Vegetação	Água	Culturas	Vegetação
Água	<b>768</b> <b>100,0%</b>	0 0,0%	0 0,0%	<b>766</b> <b>99,7%</b>	0 0,0%	2 0,3%
Culturas	0 0,0%	<b>696</b> <b>100,0%</b>	0 0,0%	0 0,0%	<b>637</b> <b>91,5%</b>	59 8,5%
Vegetação	7 1,3%	1 0,2%	<b>520</b> <b>98,5%</b>	0 0,0%	0 0,0%	<b>528</b> <b>100,0%</b>
	Acurácia geral = <b>99,6%</b>			Acurácia geral = <b>96,9%</b>		

No classificador de máxima verossimilhança gaussiana houve uma menor taxa de acerto geral, pois houve confusão entre as classes de culturas e vegetação, provavelmente provocada pela similaridade nos vetores média das duas classes e pela quebra da suposição de normalidade.

#### 4.2.2 Classificação de imagem AVIRIS

O sensor aerotransportado AVIRIS possui altíssima resolução espectral, imageando simultaneamente uma mesma cena em 220 bandas espectrais, posicionadas entre 0,38 e 2,5 micrômetros. A resolução espacial do sensor também é alta, mas depende da altura de vôo. O tamanho típico de um pixel fica em torno de 17 a 20km. A alta resolução do sensor proporciona uma riqueza de informação tão grande, que cuidados especiais são necessários ao se trabalhar com imagens AVIRIS. A escassez de amostras de treinamento pode facilmente impedir que o processo de estimação seja realizado.

A imagem utilizada, cuja verdade terrestre é conhecida, contém grande variedade de classes por tratar-se de uma zona agrícola, conforme mostra a figura 19.

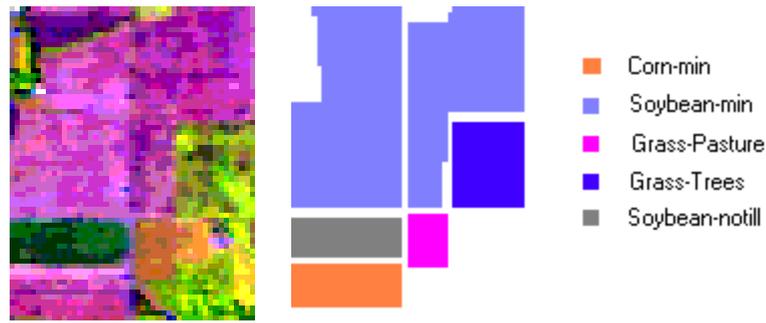


FIGURA 19: Segmento de imagem AVIRIS, composição 96-56-6 (R-G-B) e verdade terrestre

As classes espectrais presentes nesse segmento da imagem têm comportamentos espectrais bastante semelhantes, conforme ilustrado na figura 20. Foi utilizado um subconjunto de 10 bandas espectrais, extraídas de maneira sistemática.

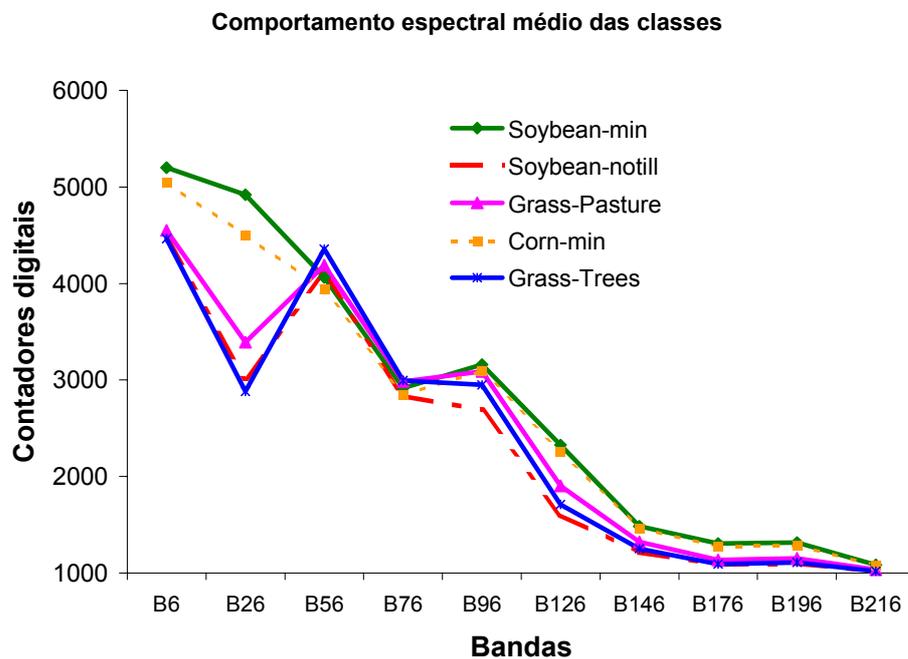


FIGURA 20: Comportamento espectral médio das classes nas dez bandas consideradas

Repetindo o procedimento de estimação utilizado na imagem Landsat-TM, chegamos a estimativas para os 44 parâmetros da discriminação logística e para

os 325 parâmetros do classificador gaussiano. Agora trabalhamos com apenas 153 amostras de teste, mas a hipótese de normalidade pode ser aceita em todas as classes e em todas as bandas espectrais. Isso sugere que a hipótese de normalidade multivariada seja verdadeira. Os resultados da classificação são apresentados na figura 21 e na tabela 10.

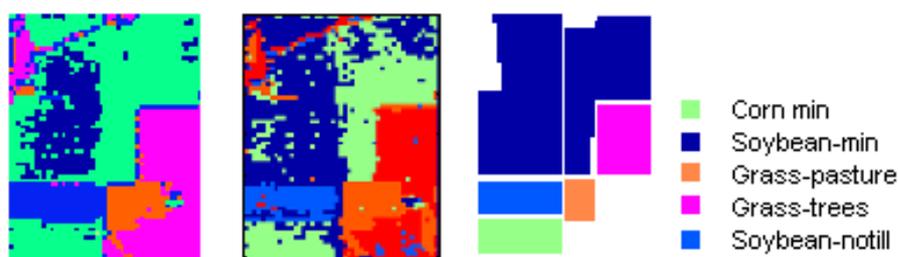


FIGURA 21: Imagens temáticas e verdade terrestre: primeira imagem classificada por máxima verossimilhança gaussiana; segunda imagem classificada por discriminação logística.

TABELA 10

Tabela de classificação da imagem AVIRIS

Classe verdadeira	Discriminação logística					Máxima verossimilhança gaussiana				
	1.	2.	3.	4.	5.	1.	2.	3.	4.	5.
1. Soybean-min	66 73,3%	0 0,0%	0 0,0%	24 26,7%	0 0,0%	58 64,4%	0 0,0%	0 0,0%	32 35,6%	0 0,0%
2. Soybean-notill	8 12,1%	53 80,3%	0 0,0%	0 0,0%	5 7,6%	0 0,0%	66 100%	0 0,0%	0 0,0%	0 0,0%
3. Grass-pasture	0 0,0%	0 0,0%	30 100%	0 0,0%	0 0,0%	0 0,0%	0 0,0%	30 100%	0 0,0%	0 0,0%
4. Corn-min	28 23,3%	0 0,0%	0 0,0%	92 76,7%	0 0,0%	8 6,7%	0 0,0%	0 0,0%	112 93,3%	0 0,0%
5. Grass-trees	0 0,0%	2 1,7%	0 0,0%	0 0,0%	118 98,3%	0 0,0%	0 0,0%	0 0,0%	0 0,0%	120 100%
	Acurácia geral = <b>84,3%</b>					Acurácia geral = <b>90,6%</b>				

Mesmo com escassez de amostras de teste, o classificador gaussiano apresentou resultados superiores. Esse fato provavelmente ocorreu devido à distribuição conjunta das bandas espectrais ter sido satisfatoriamente aproximada pela normal multivariada. O modelo logístico apresentou maior taxa de erros de classificação entre as classes *Soybean-min* e *corn-min*.

## 5 CONCLUSÕES

Não há dúvidas de que a discriminação logística consiste de uma interessante alternativa ao classificador da máxima verossimilhança gaussiana. O modelo logístico apresentou resultados equivalentes ou até melhores em situações de não-normalidade. Entretanto, sob hipótese de normalidade multivariada, o classificador gaussiano mostrou-se superior.

O fato das fronteiras de decisão da discriminação logística apresentarem-se na forma linear limita muito a separação de alguns tipos de classes. Nesse aspecto, a máxima verossimilhança gaussiana apresenta grande vantagem, pois as fronteiras de decisão têm forma elíptica.

No que se refere à estimação de parâmetros, a discriminação logística necessita obrigatoriamente da utilização de métodos numéricos para obtenção das estimativas de máxima verossimilhança, o que só é possível com a utilização de recursos computacionais. Já o método gaussiano torna-se computacionalmente mais

simples, pois utiliza diretamente as estimativas dos vetores média e matrizes covariância.

Em muitas situações reais, o número de amostras de treinamento pode não ser suficiente para obtenção de estimativas confiáveis para os parâmetros do modelo, em especial as matrizes covariância. Atualmente, a estimação de matrizes covariância é um problema crítico e que vem merecendo crescente atenção por parte de pesquisadores. Portanto, em casos de escassez de amostras de treinamento o modelo logístico pode levar vantagem por contar com um número reduzido de parâmetros. Contudo, isso não é uma regra, pois na classificação da imagem AVIRIS o fator de normalidade mostrou-se mais decisivo que a quantidade de amostras. Mesmo com um número de parâmetros superior ao número de amostras de treinamento, o classificador gaussiano foi superior.

O método da discriminação logística não se encontra implementado nos principais *softwares* de processamento de imagem, tais como *Erdas* e *Idrisi*. Provavelmente este é o motivo da falta de referencial bibliográfico sobre esse método de classificação na literatura especializada. O contrário acontece na área de estatística, onde praticamente todos os *softwares* apresentam módulos de regressão logística.

A implementação de um *software* de classificação com *interface* amigável e o estudo de alternativas, como a regressão logística quadrática, são assuntos que poderão ser estudados posteriormente.

## REFERÊNCIAS BIBLIOGRÁFICAS

- ALLISON, P. D. (1999) *Logistic Regression using the SAS System: Theory and Application*. Cary, NC: SAS Institute Inc.
- ANDERSON, J. A. (1982) Logistic Discrimination. In *Handbook of Statistics* (Vol. 2) P. R. Krishnaiah and L. Kanal (Eds.) Amsterdam: North-Holland, p. 169-191.
- BULL, S. and DONNER, A. (1987) The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of the American Statistical Association*. vol. 82, p. 1118-1122.
- COX, D.R. and SNELL, E. J. (1989). *The Analysis of Binary Data*. Second Edition. London: Chapman and Hall.
- DAY, N. and KERRIDGE, D. (1967) A general maximum likelihood discriminant. *Biometrics*, vol. 23, p. 313-324.
- EFRON, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, vol. 70, no. 352., p. 892-898.

- HAERTEL, V. and LANDGREBE, D. (1999) On the classification of classes with nearly equal spectral response in remote sensing hyperspectral image data. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 5., p. 2374-2386.
- HAUCK, W. W. and DONNER, A. (1977) Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, vol. 72, no. 360., p. 851-853.
- HOSMER, D. and LEMESHOW, S.. (1989) *Applied Logistic Regression*. New York: John Wiley & Sons.
- JAIN, A K.; DUIN R. P. and MAO, J. (2000) Statistical Pattern Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, p. 04 – 37.
- JOHNSON, R. A and WICHERN, D. W. (1982) *Applied Multivariate Statistical Analysis*. Englewood Cliffs, New Jersey: Prentice Hall.
- KRZANOWSKY, W. J. (1988) *Principles of Multivariate Analysis*. Oxford: Clarendon Press.
- McLACHLAN, G. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.
- NORUSIS, M. J. and SPSS Inc. (1999) *SPSS Regression Models 10.0*. Chicago, IL: SPSS Inc.
- PRESS, S. J. and WILSON, S. (1978) Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*. , vol. 73, p. 699-705.

RICHARDS, J. A. (1993) *Remote Sensing Digital Image Analysis: an introduction*.  
2<sup>nd</sup> ed. Berlin: Springer-Verlag.

SAS Institute Inc. (1989) *SAS/STAT User's Guide, Version 6*. 4<sup>th</sup>ed., Volume 1.  
Cary, NC: SAS Institute Inc.

SPANOS, A. (1999) *Probability Theory and Statistical Inference*. Cambridge:  
University Press.

TOU, J. and GONZALEZ, R. (1974) *Pattern Recognition Principles*. Massachusetts:  
Addison-Wesley,.

TRUETT, J. CORNFIELD, J. and KANNEL, W. (1967) A multivariate analysis of  
the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*.  
v. 20, p. 511-524.

WALKER, S. H. and DUNCAN, D. B. (1967) Estimation of the probability of an  
event as a function of several independent variables. *Biometrika* vol. 54, p. 167-  
169.