Lecture 08

LDA and QDA

https://github.com/dalcimar/MA28CP-Intro-to-Machine-Learning

Linear and Quadratic Discriminant Analysis

Are two classic classifiers, with, as their names suggest, a linear and a quadratic decision surface, respectively.

These classifiers are attractive because they have **closed-form solutions** that can be **easily computed**, are **inherently multiclass**, have proven to **work well in practice**, and have **no hyperparameters to tune**.

Both LDA and QDA can be derived from simple probabilistic models

Probabilities

The probability of discrete event A occurring is

$$P(A)$$

The continuous random variable x has a probability density function (pdf)

$$p(x)$$

For vector-valued random variables x, we write this as

$$p(\mathbf{x})$$

Conditional Probabilities

We write the conditional probability of A given B as:

$$P(A|B)$$

This means “the probability of A given B” or “given B, what is the probability of A?”

$$p(x|A)$$

Or for random variables,

$$p(\mathbf{x}|A)$$

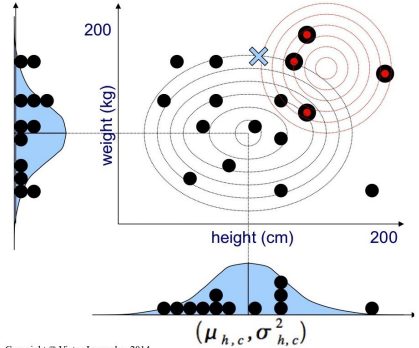
Training: Class-Conditional Probabilities

Suppose that we measure features for a large training set taken from class ω_i

Each of these training pattern has a different value **x** for the features. This can be written as the class-conditional probability:

$$p(\mathbf{x}|\omega_i)$$

In other words, how often do things in class ω_i exhibit features **x**?



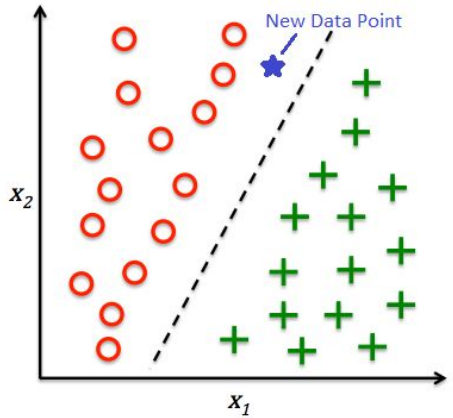
Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, …	85, …	80, …	95, …					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

Classification

When we classify, we measure the feature vector **x**, then we ask this question:

- Given that this has features **x**, what is probability that it belongs to class ω_i?

$$P(\omega_i|\mathbf{x})$$



Why We Care About Conditional Probabilities

Training gives us $p(\mathbf{x}|\omega_i)$

But we want $P(\omega_i|\mathbf{x})$

These are not the same! How are they related? **Bayes theorem!**

- Generally: $P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$

- For our purposes: $P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)\ P(\omega_i)}{p(\mathbf{x})}$

Definitions

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)\ P(\omega_i)}{p(\mathbf{x})}$$

$p(\mathbf{x}|\omega_i)$ class conditioned probability or likelihood

$P(\omega_i)$ a priori or prior probability

$p(\mathbf{x})$ evidence (usually ignored)

$P(\omega_i|\mathbf{x})$ measurement-conditioned or posterior probability

Structure of LDA and QDA classifiers

- Training:
  - Measure p(**x**|ω_i) of each class
- Prior Knowledge:
  - Measure or estimate P(ω_i) in general population
- Classification
  - Measure feature **x** for new pattern
  - Calculate posterior probabilities P(ω_i|**x**) for each class
  - Choose the one with the larger posterior P(ω_i|**x**)

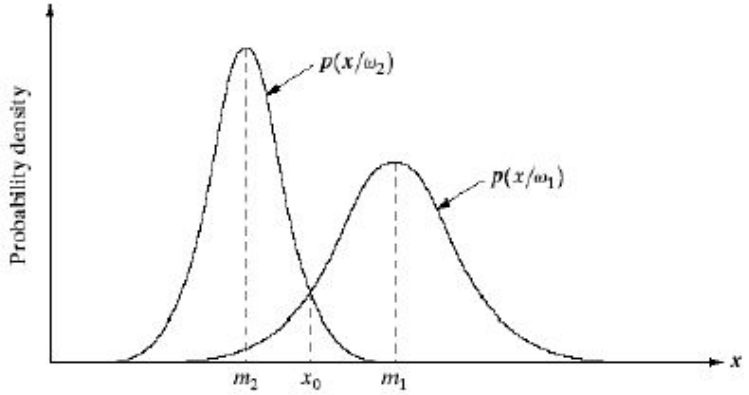
Assumption of Normality

More specifically, for **linear and quadratic discriminant analysis**, $p(\mathbf{x}|\omega_i)$ **is modeled as a multivariate Gaussian distribution with density**:

- Normally distributed class-conditional probabilities (1D):

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2}$$

FIGURE 12.10 Probability density functions for two 1-D pattern classes. The point x_0 shown is the decision boundary if the two classes are equally likely to occur.



From Probabilities to Discriminants: 1-D Case

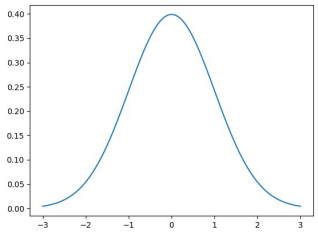
want to maximize $P(\omega_i|x) = \frac{p(x|\omega_i)\ P(\omega_i)}{p(x)}$

same as maximizing $p(x|\omega_i)\ P(\omega_i)$

which for a normal distribution is $\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2}\ P(\omega_i)$

applying logarithm 2 $\log \frac{1}{\sqrt{2\pi}} - \log \sigma_i - \frac{1}{2}(x-\mu_i)^2/\sigma_i^2 + \log P(\omega_i)$

dropping constants $\log P(\omega_i) - \log \sigma_i - \frac{1}{2}(x-\mu_i)^2/\sigma_i^2$



$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2}$$

Extending to Multiple Features

Note that the key term for a 1-D normal distribution is the **squared distance from the mean** in standard deviations

$$(x - \mu_i)^2 / \sigma_i^2$$

Can extend to multiple features by simply normalizing each feature's "distance" by the respective standard deviation, then just use minimum distance classification (remembering to use the priors as well)

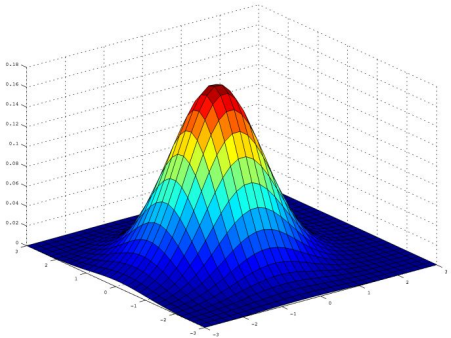
- Some call normalizing each feature by its variance naive Bayes
- So what's naive about it?
- **It ignores relationships between features**

The Multivariate Normal Distribution

In multiple dimensions, the normal distribution takes on the following form:

$$p(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{d} \frac{1}{|\mathbf{C}|^{1/2}} \; e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{T}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}$$

$$= (2\pi)^{-d/2} \, |\mathbf{C}|^{-1/2} \; e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{T}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}$$



Multivariate Gaussian distribuition

For multiple classes, each class ω_i has its own

- mean vector $\mathbf{m}_i$
- covariance matrix $\mathbf{C}_i$

The class-conditional probabilities are

$$p(\mathbf{x}|\omega_i) = (2\pi)^{-d/2} |\mathbf{C}_i|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$

From Probabilities to Discriminants: N-D Case

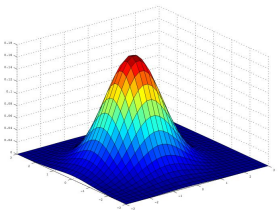
want to maximize $P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)\ P(\omega_i)}{p(\mathbf{x})}$

so maximize $p(\mathbf{x}|\omega_i)\ P(\omega_i)$

so maximize $\log p(\mathbf{x}|\omega_i) + \log P(\omega_i)$

for normal distribution $-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}_i|$
$-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log P(\omega_i)$

maximize $\log P(\omega_i) - \frac{1}{2}\log|\mathbf{C}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$



$$p(\mathbf{x}|\omega_i) = (2\pi)^{-d/2}|\mathbf{C}_i|^{-1/2}\ e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$

Mahalonobis Distance

The expression $(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}_i)$

can be thought of as $\|\mathbf{x} - \mathbf{m}_i\|^2_{\mathbf{C}^{-1}}$

- This looks like squared distance, but the inverse covariance matrix C^{-1} acts like a metric (stretching factor) on the space.
- This is the **Mahalonobis distance**!
- Pattern recognition using multivariate normal distributions is simply a minimum (Mahalonobis) distance classifier.

$$\log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

Case 1: Identity Matrix (naive Bayes)

Suppose that the covariance matrix for all classes is the identity matrix I:

$$\mathbf{C}_i = I \text{ or } \mathbf{C}_i = \sigma^2 I$$

If the data is normalized by score- z and the data is not correlated the correlation matrix is the identity matrix with unit standard deviation discriminant becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i) + \log P(\omega_i)$$

Assuming all classes are a priori equally likely

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ignoring the constant 1/2 , we can use

$$g_i(\mathbf{x}) = -(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$

Case 1: Identity Matrix

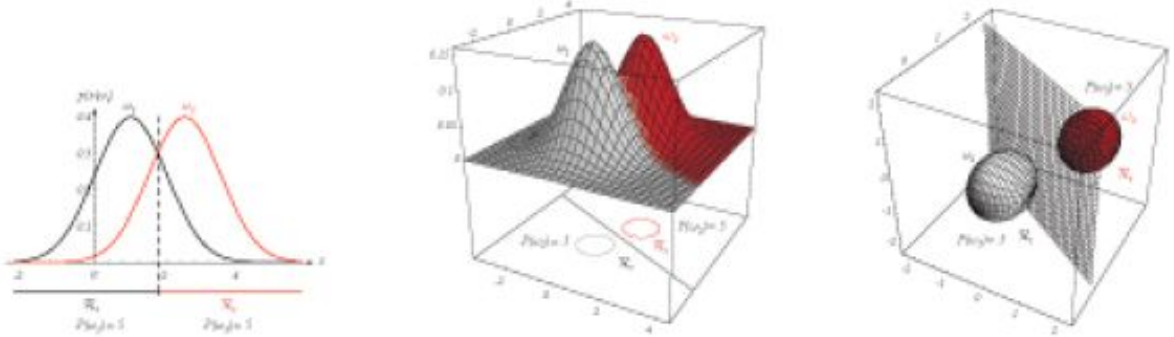


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

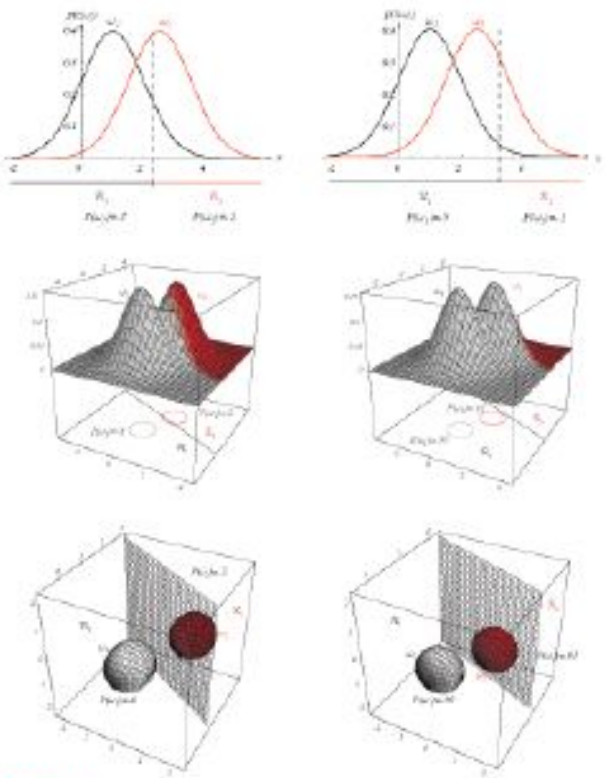


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$$\log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

Case 2: Same Covariance Matrix (LDA)

If each class has the same covariance matrix,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}(\mathbf{x}-\mathbf{m}_i) + \log P(\omega_i)$$

- Loci of constant probability are hyperellipes oriented with the eigenvectors of C:
  - eigenvectors directions of ellipse axes
  - eigenvalues variance (squared axis length) in axis directions
- The decision boundaries are still hyperplanes, though they may no longer be normal to the lines between the respective class means.

The variance-covariance matrix of three random variables X_1, X_2, X_3 is given by

$$\sum = \begin{bmatrix} 1 & 0{\cdot}63 & 0{\cdot}4 \\ 0{\cdot}63 & 1 & 0{\cdot}35 \\ 0{\cdot}4 & 0{\cdot}35 & 1 \end{bmatrix}.$$

Write its factor model.

Case 2: Same Covariance Matrix

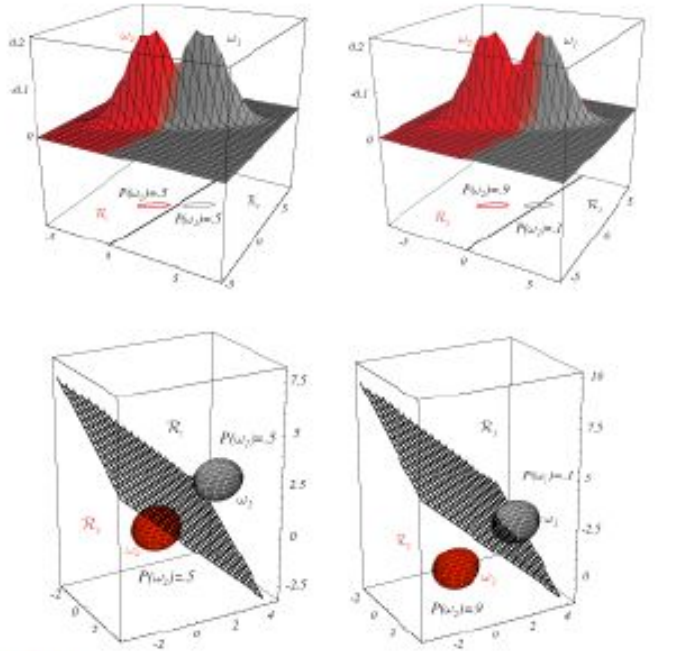


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$$\log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

Case 3: Different Covariances for Each Class (QDA)

Suppose that each class has its own arbitrary covariance matrix (the most general case):

$$\mathbf{C}_i \neq \mathbf{C}_j$$

$$g_i(\mathbf{x}) = \log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

- Loci of constant probability for each class are hyper ellipse oriented with the eigenvectors of C_i for that class.
- Decision boundaries are quadratic, specifically, hyper ellipses or hyper hyperboloids.

$$\mathrm{Var}[X] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Case 3: Different Covariances for Each Class

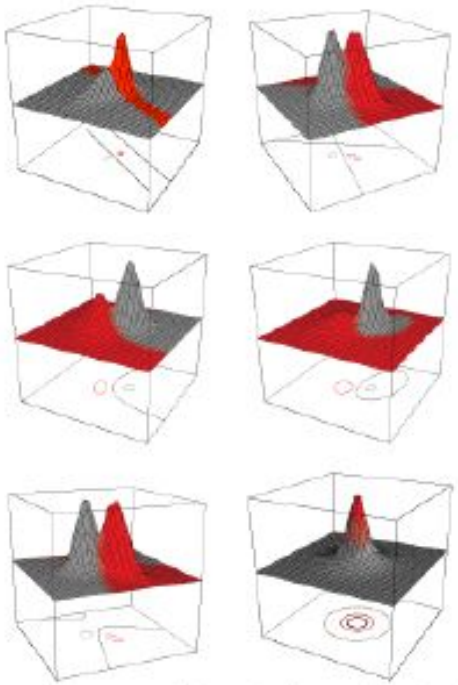
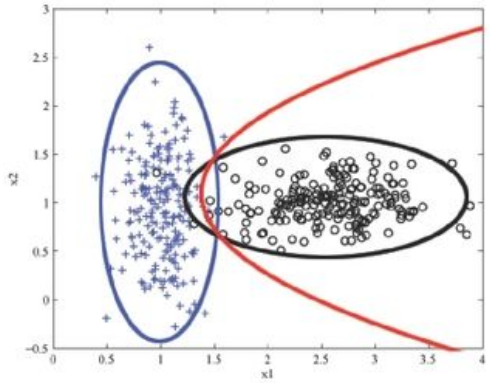
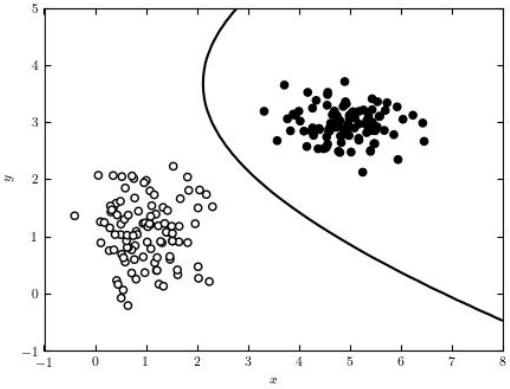
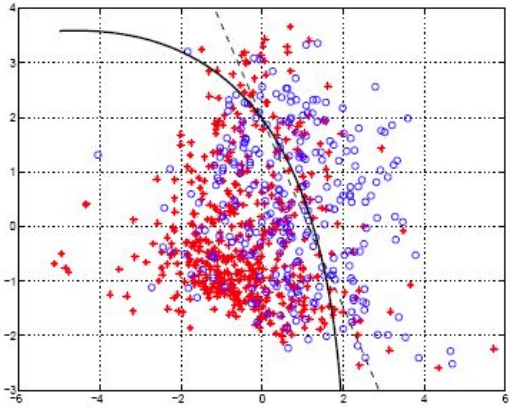


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

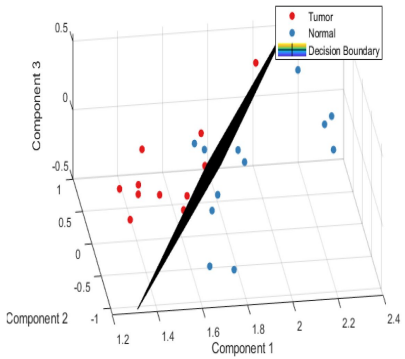




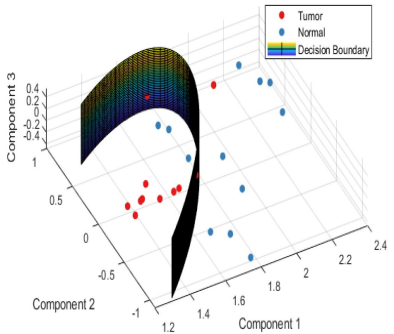


Case 3: Different Covariances for Each Class

3D Case



(a)



(b)

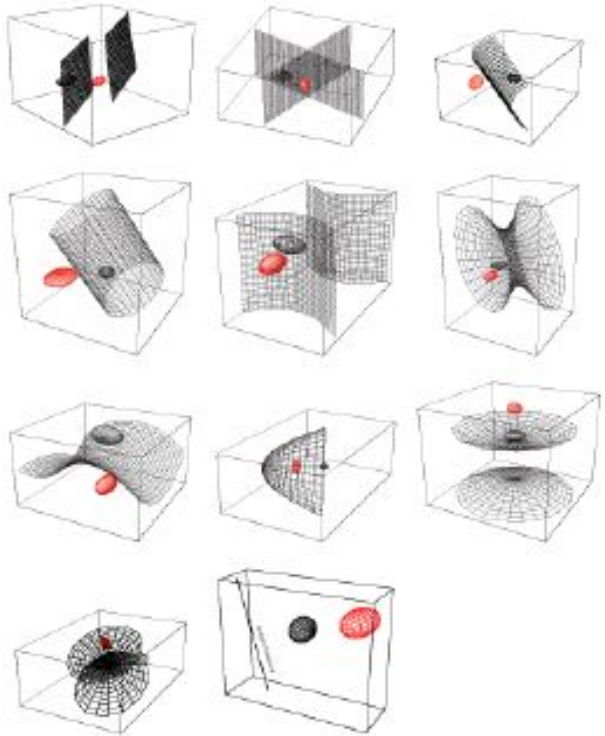


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.

Case 3: Different Covariances for Each Class

Multiclass

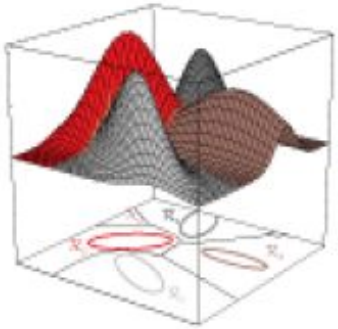
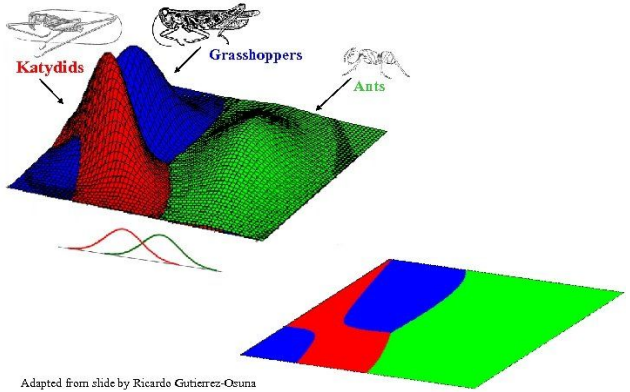


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Naïve Bayesian Classifier has a piecewise quadratic decision boundary



Adapted from slide by Ricardo Gutierrez-Osuna

LDA x Fisher linear discriminant

Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant

- The terms **Fisher's linear discriminant** and **LDA** are often **used interchangeably**, although Fisher's original article actually describes a **slightly different discriminant**, which does **not make some of the assumptions of LDA** such as **normally distributed classes** or **equal class covariances**.

Exemplo numérico

- Treinamento
  - Determinar médias de matriz de covariâncias

$$\mathbf{X} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \\ 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

$$C = \begin{pmatrix} 0.2403 & -0.2694 \\ -0.2694 & 1.9742 \end{pmatrix}$$

$$C^{-1} = \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} 2.88 & 5.6771 \end{bmatrix} \qquad g = 2$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}(\mathbf{x}-\mathbf{m}_i) + \log P(\omega_i)$$

Exemplo numérico

Classificação: $\mathbf{x} = [3 \quad 7]$ $\qquad P(i \mid \mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - 2\ln(P(i))$

- Classe 1: $P(1 \mid \mathbf{x}) = ([3 \quad 7]-[3.05 \quad 6.38])^T \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix} ([3 \quad 7]-[3.05 \quad 6.38]) + 1.1192$

  $\boldsymbol{\mu}_1 = [3.05 \quad 6.38]$ $\qquad P(1) = 4/7$

- Classe2: $P(2 \mid \mathbf{x}) = ([3 \quad 7]-[2.67 \quad 4.73])^T \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix} ([3 \quad 7]-[2.67 \quad 4.73]) + 1.6946$

  $\boldsymbol{\mu}_2 = [2.67 \quad 4.73]$ $\qquad P(2) = 3/7$

$$g_i(\mathbf{x}) = \log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

Exemplo numérico

$$\mathbf{X}_{classe1} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix} \qquad C_1 = \begin{pmatrix} 0.1876 & -0.4127 \\ -0.4127 & 1.1372 \end{pmatrix} \qquad C_1^{-1} = \begin{pmatrix} 26.3961 & 9.5785 \\ 9.5785 & 4.3552 \end{pmatrix}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3.05 & 6.38 \end{bmatrix} \qquad P(1) = 4/7$$

$$\mathbf{X}_{classe2} = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} \qquad C_2 = \begin{pmatrix} 0.3141 & -0.7308 \\ -0.7308 & 1.8785 \end{pmatrix} \qquad C_2^{-1} = \begin{pmatrix} 33.5580 & 13.0550 \\ 13.0550 & 5.6111 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 2.67 & 4.73 \end{bmatrix} \qquad P(2) = 3/7$$

Exemplo numérico

- QDA: P(1|**x**)=-0.8791; P(2|**x**)=50.9385
  - x é da **classe 2**

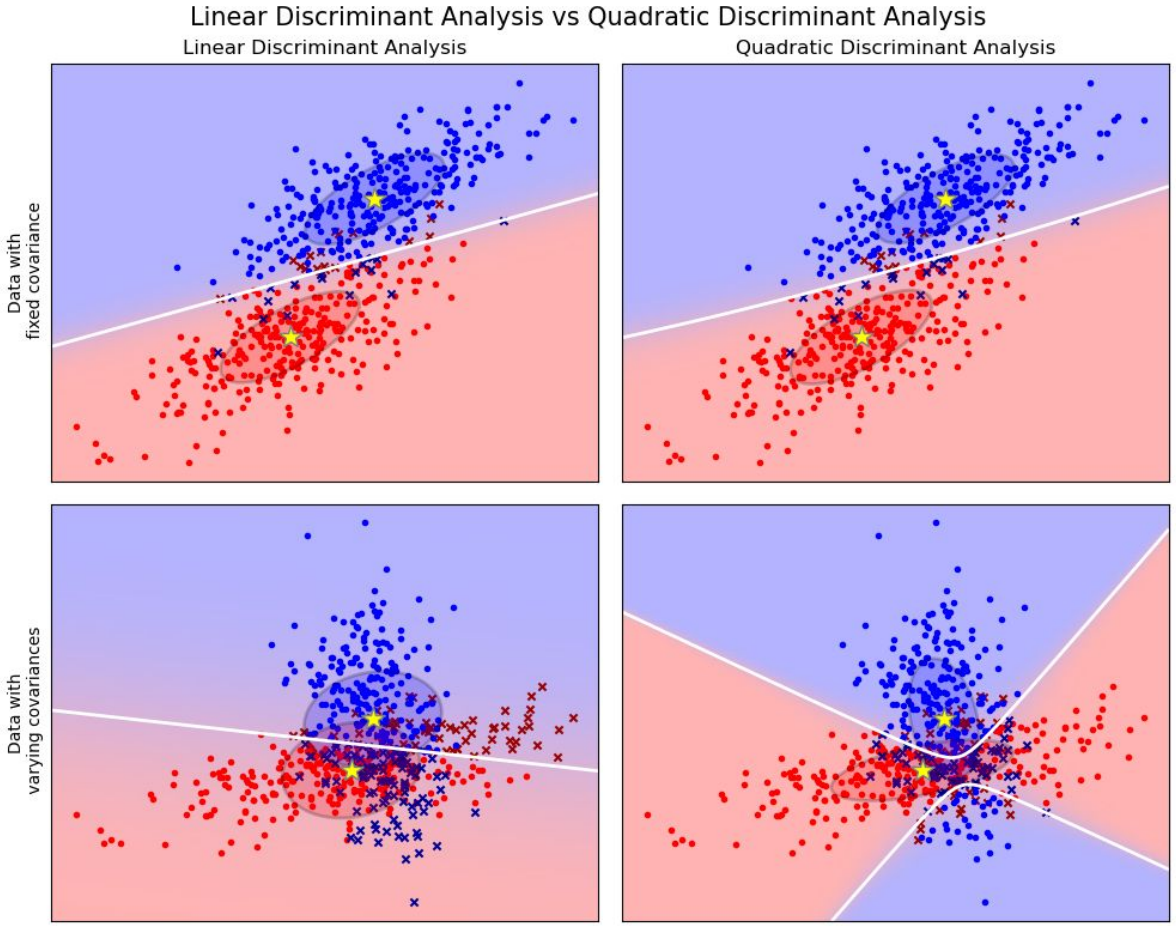
$$g_i(\mathbf{x}) = \log P(\omega_i) - \tfrac{1}{2}\log|\mathbf{C}_i| - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)$$

- Fisher: P(1|**x**)=1.3198; P(2|**x**)=6.3157
  - x é da **classe 2**

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}(\mathbf{x}-\mathbf{m}_i) + \log P(\omega_i)$$

- Bayes: P(1|**x**)=1.5061; P(2|**x**)=6.9564
  - x é da **classe 2**

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$


Linear Discriminant Analysis vs Quadratic Discriminant Analysis
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances

Conclusões

Atributos

- Necessariamente numéricos
  - matriz de correlação e médias
- Necessariamente simétricos
  - Discretos / Continuos
- Suporta probabilidades a priori
- Assume que atributos são igualmente importantes
- Seleção de atributos
  - Wrapper ou Embedded

Conclusões

- Capaz de classificar padrões com valores ausentes
  - Matriz de correlação diferente
- Robusto a ruídos isolados
  - Afeta pouco as probabilidades
- Robusto a atributos irrelevantes
  - Afeta pouco as probabilidades da classe
- Complexidade computacional
  - $O(n^3)$ e $O(n)$

Conclusões

- Hipótese de dependência entre atributos
- Determinístico
- Não parametrizado
- Melhor que o Naive Bayes, especialmente com atributos correlacionados