Bayesian Classification

CS 650: Computer Vision

Bryan S. Morse
BYU Computer Science

Training: Class-Conditional Probabilities

- Suppose that we measure features for a large training set taken from class ω_i.
- Each of these training patterns has a different value $\mathbf{x}$ 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 $\mathbf{x}$?

Classification

When we classify, we measure the feature vector $\mathbf{x}$, then we ask this question:

"Given that this has features $\mathbf{x}$, what is the probability that it belongs to class ω_i?".

Mathematically, this is written as

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

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 (Revisited)

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 a Bayesian Classifier

Training:
Measure $p(\mathbf{x}|\omega_i)$ for each class.

Prior Knowledge:
Measure or estimate $P(\omega_i)$ in the general population.
(Can sometimes aggregate the training set if it is a reasonable sampling of the population.)

Classification:

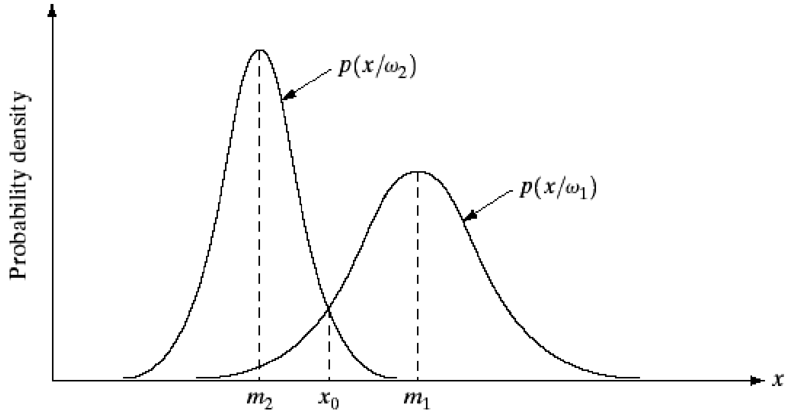
1. Measure feature ($\mathbf{x}$) for new pattern.
2. Calculate posterior probabilities $P(\omega_i|\mathbf{x})$ for each class.
3. Choose the one with the larger posterior $P(\omega_i|\mathbf{x})$.

Example

Normally distributed class-conditional probabilities:

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

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	$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)\ \ P(\omega_i)}{p(x)}$
same as maximizing	$p(x \mid \omega_i)\ \ P(\omega_i)$
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	$\log \frac{1}{\sqrt{2\pi}} - \log \sigma_i - \frac{1}{2}(x-\mu_i)^2/\sigma_i^2 + \log P(\omega_i)$
dropping constants	$\log P(\omega_i) - \log \sigma_i - \frac{1}{2}(x-\mu_i)^2/\sigma_i^2$

Extending to Multiple Features

- Note that the key term for a 1-D normal distribution is

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

  the squared distance from the mean *in standard deviations*
- 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)

Extending to Multiple Features

- 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:

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

[See examples in Mathematica]

Multivariate Normal Bayesian Classification

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

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

From Probabilities to Discriminants

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

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 $\mathbf{C}^{-1}$ 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.

Case 1: Identity Matrix

Suppose that the covariance matrix for all classes is the identity matrix I:
$$\mathbf{C}_i = I \text{ or } \mathbf{C}_i = \sigma^2 I$$

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 ω_i are *a priori* equally likely,
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$

Ignoring the constant $\frac{1}{2}$, we can use
$$g_i(\mathbf{x}) = -(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$

Example: Equal Priors

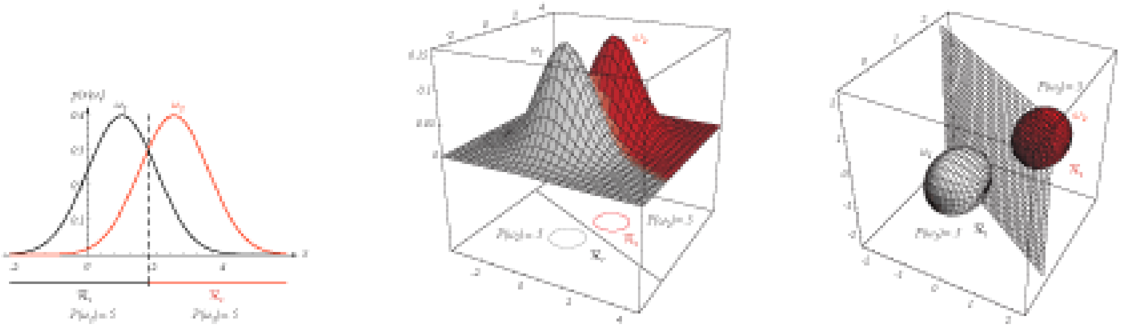


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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.

Examples: Different Priors

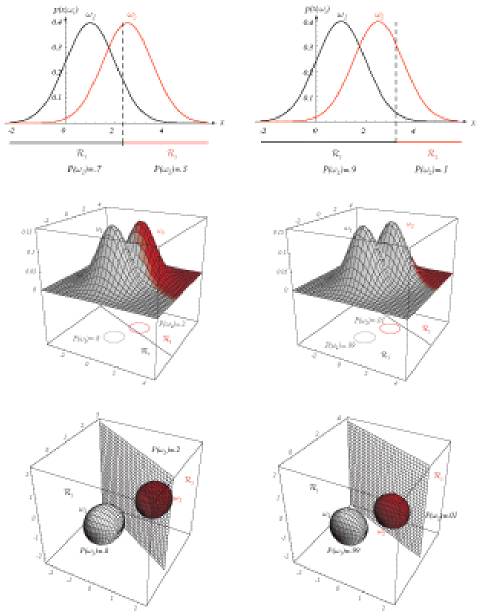


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Case 2: Same Covariance Matrix

- 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.

Examples

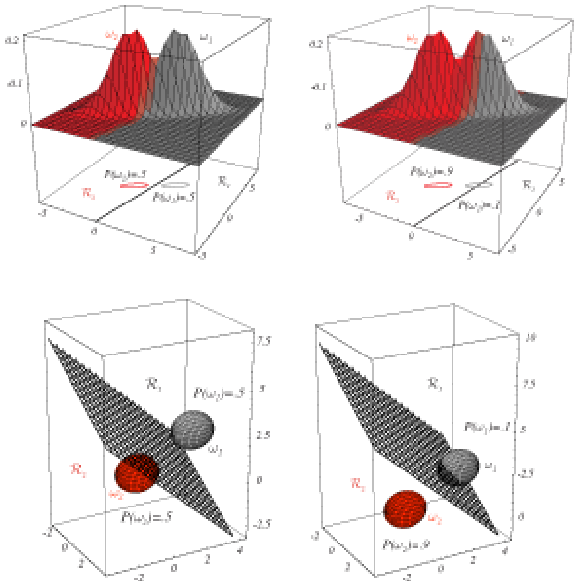


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Case 3: Different Covariances for Each Class

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

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

- Loci of constant probability for each class are hyperellipes oriented with the eigenvectors of $\mathbf{C}_i$ for that class.
- Decision boundaries are quadratic, specifically, *hyperellipses* or *hyperhyperboloids*.

[See examples in Mathematica]

Examples: 2-D

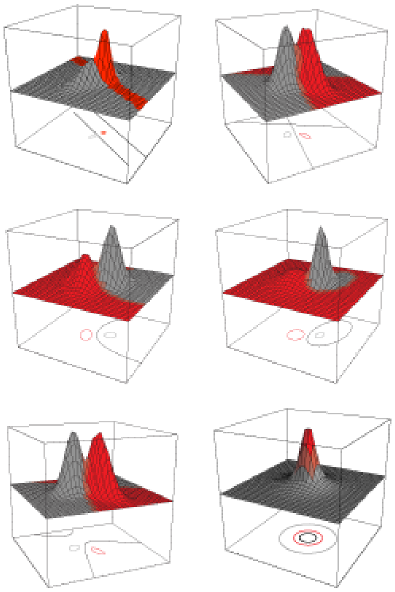


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Examples: 3-D

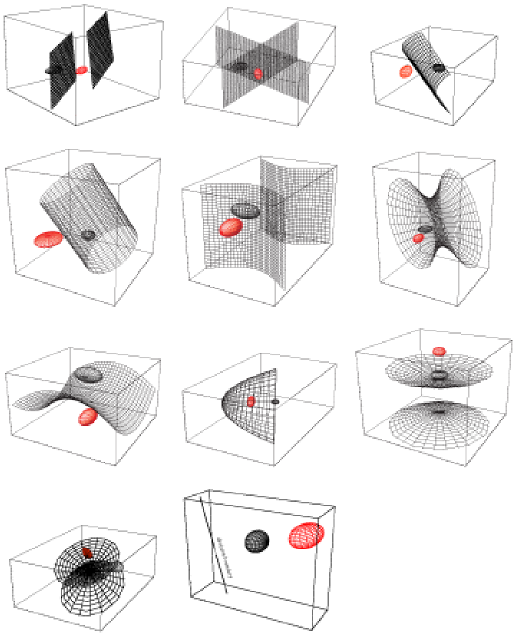


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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

Example: Multiple Classes

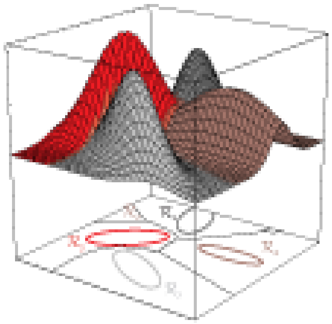


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.