University of Cambridge

MPhil in Computer Speech Text &
Internet Technology

Module: Speech Processing 1

Lecture 4: Bayes’ Decision Rule

Michaelmas 2002
Bayes’ Decision Rule

The overall aim of the classifier may be summed up as: “How do we build a pattern classifier that minimises the average probability of error?”

Consider a system where

- The feature vector is of dimension $d$, and there are $K$ classes, denoted by the symbols $\omega_1, \omega_2, \ldots, \omega_K$.
- Suppose we have an idea of how frequent each class is; this information is modelled as the *a priori* probabilities, or prior distribution, $P(\omega_1), P(\omega_2), \ldots, P(\omega_K)$.
- For each class using the observations, denoted as $\mathbf{x}$, we train a *class condition* probability density function. This is the probability density function in the $d$ dimensional space of features of the data for a particular class, sometimes referred to as the *likelihood*. We denote the likelihood function for class $k$ as $p(\mathbf{x}|\omega_k)$.

Bayes’ rule lets us combine the prior probabilities and the likelihood functions to calculate a *posterior* probability of class membership.

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{\sum_{k=1}^{K} p(\mathbf{x}|\omega_k) P(\omega_k)}, \quad j = 1, 2, \ldots, K$$
Bayesian Decision Theory

The goal in creating a decision rule here is to minimise the average probability of error.

\[ P(\text{error}) = \int p(\text{error}, x)dx \]
\[ = \int P(\text{error}|x)p(x)dx \]

For a two class problem, the conditional probability of error, (i.e. the error probability, given a value for the feature vector), can be written as

\[ P(\text{error}|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases} \]

A decision rule minimises this conditional probability of error, applied to every example, will yield the rule that minimises the average probability of error. This leads to Bayes’ decision rule, which for a two class problem is

\[ \text{Decide } \begin{cases} \text{Class } \omega_1 & \text{if } P(\omega_1 | x) > P(\omega_2 | x) \\ \text{Class } \omega_2 & \text{otherwise} \end{cases} \]

The two-class decision rule divides the complete space into two regions, decide class \( \omega_1 \) in region \( \mathcal{R}_1 \) and \( \omega_2 \) in \( \mathcal{R}_2 \). The probability of error is then

\[ P(\text{error}) = P(x \in \mathcal{R}_2, \omega_1) + P(x \in \mathcal{R}_1, \omega_2) \]
\[ = P(x \in \mathcal{R}_2 | \omega_1)P(\omega_1) + P(x \in \mathcal{R}_1 | \omega_2)P(\omega_2) \]
\[ = \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)dx + \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)dx \]

Bayes’ Decision Theory (cont)

The error regions for a two-class problem are shown below (from DHS). The decision boundary $x^*$ is set to $x_B$ for minimum error.

For the two-class case the Bayes’ minimum average error decision rule could be written as a ratio of posterior probabilities:

$$\frac{P(\omega_1 \mid x)}{P(\omega_2 \mid x)} \begin{cases} \omega_1 > 1 \\ \omega_2 \end{cases}$$

For multi-class problems, we calculate all the $K$ posterior probabilities. The minimum error classifier is obtained from

$$\max_j p(x \mid \omega_j) P(\omega_j)$$
Decision Boundaries

*Decision boundaries* partition the feature space into regions with a class label associated with each region. Initially only two class problems will be considered.

We would like the decision boundary to reflect the Bayes’ decision rule from the previous slide. For the two class problem it is clear(?) that the decision boundary will occur when the posterior probability of class 1 and class 2 are equal. This tells us that the decision occurs when the posterior probability is

\[
P(\omega_1|x) = \frac{1}{2} = \frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)}
\]

We therefore say that a point \( x \) is on the decision boundary when it satisfies

\[
p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2)
\]

We shall see that the nature of the decision boundary will depend on the form of the class-conditional densities and priors used.

This lecture will examine the nature of the decision boundary for a particular classifier using Gaussian distributions to model the class-conditional density distributions.

But first a couple of simple examples ...
Simple Example

A lecturer needs to classify two groups of people the “early” risers (class 1, $\omega_1$) and the “late” risers (class 2, $\omega_2$). There are equal numbers in both groups. The lecturer observes at what time ($x$) during the hour lecture they arrive in the class. From previous observations he knows that the class-
conditional probability density functions describing the arrival times are

$$p(x|\omega_1) = 2(1 - x); \quad p(x|\omega_2) = 2x; \quad 0 \leq x \leq 1$$

$$0; \quad \text{otherwise}$$

At what time should the lecturer start classifying students as late risers? What percentage of students are expected to be misclassified?
Example (cont)

From the lecturer’s prior knowledge of the class he knows that the prior probabilities for both classes are equal, \( P(\omega_1) = 0.5 \) and \( P(\omega_2) = 0.5 \).

Using Bayes rule the class-conditional densities and prior probabilities may be transformed into class posteriors. Thus

\[
P(\omega_1|x) = \frac{0.5(2(1 - x))}{0.5(2(1 - x)) + 0.5(2x)} = (1 - x)
\]

Plotting the posterior probability for class 1 against \( x \)

Where should the decision boundary be placed? What is the probability of error?
Effect of Priors

The following year the lecturer is informed that there are twice as many early risers in the class than late risers. So the lecturer updates his prior

\[ P(\omega_1) = \frac{2}{3}; \quad P(\omega_2) = \frac{1}{3} \]

The posterior distributions now look like

Increasing the prior on class 1 has shifted the decision boundary to the right and made the posterior distribution non-linear.

Where should the decision boundary be placed?
Again you should be able to derive the expression for this line.
Multivariate Gaussian Distributions

One very common PDF used in pattern processing is the multivariate Gaussian (or Normal) distribution. For $d$-dimensional data the class conditional form (expressed for class $\omega_1$) is

$$p(x|\omega_1) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right)$$

If the covariance matrix is diagonal this expression may be simplified to

$$p(x|\omega_1) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi \sigma_i^2}} \exp \left( -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right)$$
Number of Model Parameters

The number of parameters is very important when trying to ensure robust (accurate) parameter estimation.

1. **Mean**: The number of parameters is the dimensionality of the data, $d$.

2. **Covariance Matrix**: Two forms of covariance matrix are commonly used

   (a) **Diagonal**: Here the elements of the feature vector are assumed *uncorrelated*.

   $$
   \Sigma = \begin{bmatrix}
   \sigma_1^2 & 0 & \ldots & 0 \\
   0 & \sigma_2^2 & \ldots & 0 \\
   \vdots & \vdots & \ddots & \vdots \\
   0 & \ldots & \ldots & \sigma_d^2 \\
   \end{bmatrix}
   $$

   The number of parameters is again the dimensionality of the data, $d$.

   (b) **Full**: Elements may be correlated. A *symmetric* matrix is used. Number of parameters is $\frac{d(d+1)}{2}$.

   $$
   \Sigma = \begin{bmatrix}
   \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1d} \\
   \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2d} \\
   \vdots & \vdots & \ddots & \vdots \\
   \sigma_{1d} & \ldots & \ldots & \sigma_d^2 \\
   \end{bmatrix}
   $$
Central Limit Theorem

One of the reasons for the popularity of the Gaussian distribution is the central limit theorem. Consider a set of independent samples, $x_1, \ldots, x_n$. Take the average of all these samples

$$
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
$$

Irrespective of how the original samples were distributed the distribution of the sum, $\bar{x}$, in the limit as $n \to \infty$ is Gaussian.

A simple example is when each sample is uniformly distributed between 0 and 1. The next slide shows how the distribution varies as we increase the number of points.

As the number of variables added together increases the distribution becomes more Gaussian. Note that the mean is constant (as expected) at about 0.5.
Central Limit Example

All variables are uniformly distributed between 0 and 1, and are independent of one another.
Gaussian Decision Boundaries

Assume we have a 2-class problem, $\omega_1$ and $\omega_2$. The class-conditional PDFs, $p(x|\omega_1)$ and $p(x|\omega_2)$, are both Gaussian. The priors, $P(\omega_1)$ and $P(\omega_2)$, are known. What do the decision boundaries look like?

Consider three distinct cases

1. $\Sigma_i = \sigma^2 I$ : all the class covariance matrices are scaled versions of the identity matrix.
2. $\Sigma_i = \Sigma$ : all the class covariances are the same.
3. $\Sigma_i$ arbitrary : no restrictions on the form of covariance matrix.

Depending on the restrictions the nature of the decision boundary will vary.

Irrespective of the form of the distribution a point $x$ on the decision boundary satisfies

$$p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2)$$

or taking logs

$$\log(p(x|\omega_1)) + \log(P(\omega_1)) = \log(p(x|\omega_2)) + \log(P(\omega_2))$$

We need to find the expression in terms of $\mu_1, \mu_2, \Sigma_1, \Sigma_2, P(\omega_1)$ and $P(\omega_2)$ that must be satisfied for all points, $x$, that lie on the decision boundary.
**Special Case 1: \( \Sigma_i = \sigma^2 \mathbf{I} \)**

This is the case where the covariance matrices of both classes are equal, diagonal and the variances of all features are the same. In two dimensions, the scatter of the two classes would look circular.

The discriminant function in this case can be reduced to a simple expression. The determinant and inverse of the covariance matrix can be reduced to very simple expressions. \( |\Sigma_i| = \sigma^{2d} \) and \( \Sigma_i^{-1} = \frac{1}{\sigma^2} \mathbf{I} \).

Hence equating the two expressions to find the position of the decision boundary

\[
\log \left( \frac{P(\omega_1)}{\sqrt{\sigma^{2d}(2\pi)^d}} \right) + \frac{1}{\sigma^2} \left( -\frac{1}{2}(\mathbf{x} - \mu_1)'(\mathbf{x} - \mu_1) \right) = \\
\log \left( \frac{P(\omega_2)}{\sqrt{\sigma^{2d}(2\pi)^d}} \right) + \frac{1}{\sigma^2} \left( -\frac{1}{2}(\mathbf{x} - \mu_2)'(\mathbf{x} - \mu_2) \right)
\]

Cancelling terms yields it is simple to see that

\[
\frac{1}{\sigma^2}(\mu_1 - \mu_2)'\mathbf{x} = \frac{1}{2\sigma^2}(\mu_1'\mu_1 - \mu_2'\mu_2) + \log \left( \frac{P(\omega_2)}{P(\omega_1)} \right)
\]

Since the means are fixed (after training!), this may be written as

\[
\mathbf{w}'\mathbf{x} = b
\]

Which is the equation of a line.
Example

Consider the case of Gaussian distributions with identity matrix covariance matrices $\Sigma_i = I$ and means at -1,-1 and 1,1. The priors on the two classes are equal. So we have

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma_1 = \Sigma_2 = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Hence for previous slide

$$w = \frac{1}{1} \left( \begin{bmatrix} -1 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$

You should be able to find $b$. The lines of equal likelihood and the decision boundary are:
Special Case 2: $\Sigma_i = \Sigma$

In this case we have full but equal covariance matrices for the two classes.

$$\log(p(x|\omega_1)) = -\frac{1}{2}(x - \mu_1)'\Sigma^{-1}(x - \mu_1) + \log \left( \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \right)$$

The final term is a constant (i.e. independent of $x$). This classifier is said to compute the Mahalanobis distance, given by the expression

$$(x - \mu_1)'\Sigma^{-1}(x - \mu_1).$$

When compared to the Euclidean distance in the previous case, this is a weighted distance, where different features get different weights attached to them.

Let's now derive an expression for the class boundary in the input space. Again cancelling terms from either side

$$(\mu_1 - \mu_2)'\Sigma^{-1}x = \frac{1}{2}(\mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2) + \log \left( \frac{P(\omega_2)}{P(\omega_1)} \right)$$

The equation of this class boundary is again of the form

$$w'x = b$$
General Case

For the general case where the two covariance matrices are not equal the form of the decision boundary is quadratic

\[ x'Ax + b'x + c = 0 \]

At this stage you should be able to derive this expression.

If you cannot solve this expression ask me next time!

You should find that

\[ A = \Sigma_1^{-1} - \Sigma_2^{-1} \]
\[ b' = 2 \left( \mu_2'\Sigma_2^{-1} - \mu_1'\Sigma_1^{-1} \right) \]

and the constant is given by

\[ c = \mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2 + 2\ln\left( \frac{P(\omega_2)|\Sigma_1|^{1/2}}{P(\omega_1)|\Sigma_2|^{1/2}} \right) \]
Example Decision Boundaries

Arbitrary Gaussian distributions can lead to general hyper-quadratic boundaries. The following figures (from DHS) indicate this. Note that the boundaries can of course be straight lines and the regions may not be simply connected.
Convolution Solution

From the question in lecture 3, we want to find $p_z(z)$. From notes the distribution of the sum involves convolution

$$p_x(x) * p_y(y) = \int_{-\infty}^{\infty} p_x(x)p_y(z - x)dx$$

For this problem $p_x(x)$ and $p_y(y)$ are both uniform distributions from 0 to 1 and

$$z = 20(x + y + 1)$$

Just consider the distribution of $z' = x + y$. We need to calculate the correct limits. For the product in the integral to be non-zero both these inequalities must be satisfied

$$0 \leq x \leq 1$$
$$0 \leq z' - x \leq 1$$

The second bound may be rewritten as

$$z' - 1 \leq x \leq z'$$

If $z'$ lies in the range 0 to 1

$$\int_{-\infty}^{\infty} p_x(x)p_y(z' - x)dx = \int_{0}^{z'} p_x(x)p_y(z' - x)dx = z'$$

If $z'$ lies in the range 1 to 2

$$\int_{-\infty}^{\infty} p_x(x)p_y(z' - x)dx = \int_{z' - 1}^{1} p_x(x)p_y(z' - x)dx = 2 - z'$$
Bayes' Decision Rule

**Convolution Solution (cont)**

We want the distribution of $z$, where

$$z = 20(z' + 1)$$

The PDF of $z$ is therefore (scaling?)

$$p_z(z) = \begin{cases} 
\frac{1}{20^2}(z - 20), & 20 \leq z \leq 40 \\
\frac{1}{20^2}(60 - z), & 40 \leq z \leq 60 \\
0, & \text{otherwise}
\end{cases}$$

Check the expected value

$$\mathbb{E}\{z\} = \frac{1}{20^2} \int_{20}^{40} z(z - 20) \, dz + \frac{1}{20^2} \int_{40}^{60} z(60 - z) \, dz$$

$$= \frac{50}{3} + \frac{70}{3} = 40$$