CSIS8502

2. Bayes Decision Theory

Bayes Decision Theory

Assumptions:

- The decision problem is posed in probabilistic terms
- All the relevant probability values (or distributions) are known.

Let ω denotes the state of nature (i.e. pattern classes)

$$\text{e.g.} \quad \omega = \begin{cases} \omega_1 & \text{for salmon} \\ \omega_2 & \text{for sea bass} \end{cases}$$

There exist some *a priori* probability $P(\omega_1)$ and $P(\omega_2)$, i.e. before any observation is done.
e.g. If 2/3 of all the fishes are salmon and 1/3 are sea basses, then we know that

$$\begin{aligned} P(\omega_1) &= 2/3, \quad \text{and} \\ P(\omega_2) &= 1/3 \end{aligned}$$

without any observation.

Note

1. for a 2-class case, $P(\omega_1) + P(\omega_2) = 1$, and $P(\omega_1) \geq 0$, $P(\omega_2) \geq 0$.
2. Capital P is used for probability value, while small p is used for probability density function (when input is continuous).

To recognize the pattern, we need to make observation on the object. The observation is a measurement $\vec{x}$, a feature vector representing the pattern.
The problem becomes:

- Given $\vec{x}$, i.e. Observing the length and lightness, how to make decision?
- How to represent information (in numerical values) such as:
  - salmon is longer than sea bass
  - salmon is lighter than sea bass, etc.

State-conditional probability density

- $p(\vec{x}|\omega_j)$ the probability density function for $\vec{x}$ given that the state of nature is ω_j (i.e. The fish is salmon (ω_1) or sea bass (ω_2)).
- These probability density function can be obtained by observing a large number of pattern samples (statistical).
- e.g. By measuring a large number of samples of salmon (or sea bass), we can estimate the densities for both length and lightness, given that the class is salmon (or sea bass).

State-conditional probability density

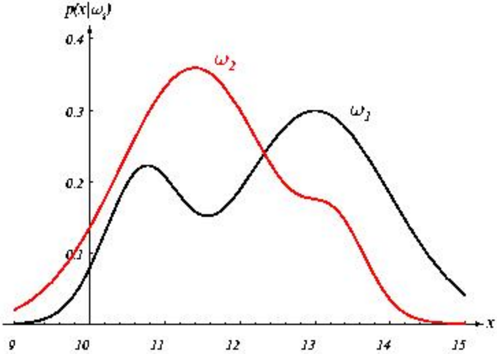

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Rule

- Suppose the *a priori* probability $P(\omega_j)$ and the state-conditional densities $p(\vec{x}|\omega_j)$ are known.
- we want to compute the *a posteriori* probability $P(\omega_j|\vec{x})$, i.e. Given the observation $\vec{x}$, we want to know the probability of it being a salmon or sea bass.
- We can use Bayes rule

$$P(\omega_j|\vec{x}) = \frac{p(\vec{x}|\omega_j)P(\omega_j)}{p(\vec{x})},$$

where

$$p(x) = \sum_{j=1}^{2} p(\vec{x}|\omega_j)P(\omega_j)$$

Example

For example, given the density function as in fig. 2.1, we can obtain the *a posteriori* probability, for $P(\omega_1) = \frac{2}{3}$, $P(\omega_2) = \frac{1}{3}$.

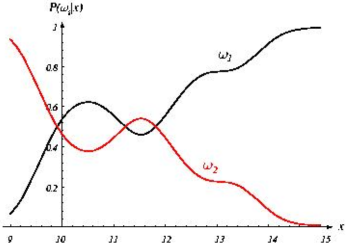

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

For more than 2 classes, just replace 2 by c, where c is the number of classes.

The Bayes Decision Rule

- If we have an observation such that $P(\omega_1|\vec{x})$ is greater than $P(\omega_2|\vec{x})$ we will incline to decide the object is ω_1.
- The probability of error is then:

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

- The average prob. of error is given by:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, \vec{x}) d\vec{x} = \int_{-\infty}^{\infty} P(\text{error}|\vec{x}) p(\vec{x}) d\vec{x}$$

- Consider the Bayes decision Rule:
  decide ω_1 if $P(\omega_1|\vec{x}) > P(\omega_2|\vec{x})$; otherwise decide ω_2.
- Under this rule,

$$P(\text{error}|\vec{x}) = \min\left[P(\omega_1|\vec{x}), P(\omega_2|\vec{x})\right]$$

- Hence we can see that $P(\text{error})$ will be minimized.

Discriminant Functions

- The classifier is viewed as a machine that computes c discriminant functions $g_i(\vec{x}), \ \ i = 1, \cdots, c$.
- The classifier selects class ω_i if

$$g_i(\vec{x}) > g_j(\vec{x}) \ \ \forall j \neq i$$

- fig 2.5 shows such machine:

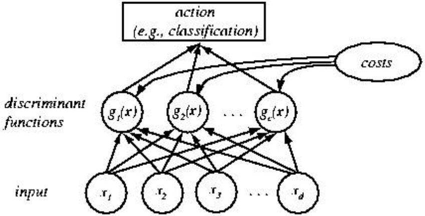

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes *d* inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Minimum Error Rate Classification

- For minimum error rate classification,

$$g_i(\vec{x}) = P(\omega_i|\vec{x})$$

- We can replace $g_i(\vec{x})$ by $f(g_i(\vec{x}))$ where f is a monotonically increasing function without changing the result.

Example

$g_i(\vec{x}) = P(\omega_i|\vec{x})$ for minimum-error rate, since

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

the following classifiers are equivalent:

$$\begin{aligned} g_i(\vec{x}) &= \frac{p(\vec{x}|\omega_i)P(\omega_i)}{p(\vec{x})} \\ g_i(\vec{x}) &= p(\vec{x}|\omega_i)P(\omega_i) \\ g_i(\vec{x}) &= \log p(\vec{x}|\omega_i) + \log P(\omega_i) \end{aligned}$$

Decision Regions

- The effect of any decision rule is to divide the feature space into c decision regions $\Re_1, \cdots, \Re_c$.
- If $g_i(\vec{x}) > g_j(\vec{x}) \; \forall j \neq i$, then $\vec{x}$ is in $\Re_i$.
- If $\Re_i$ and $\Re_j$ are contiguous, the equation for the decision boundary separating them is

$$g_i(\vec{x}) = g_j(\vec{x})$$

Example (2 category case)

- one discriminant function can be used, instead of 2.
  Let $g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x})$
  Decide ω_1 if $g(x) > 0$, otherwise decide ω_2.
- For minimum error rate,

$$\begin{aligned} g(\vec{x}) &= P(\omega_1|\vec{x}) - P(\omega_2|\vec{x}), \text{ or} \\ g(\vec{x}) &= \log\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} + \log\frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

2-class case

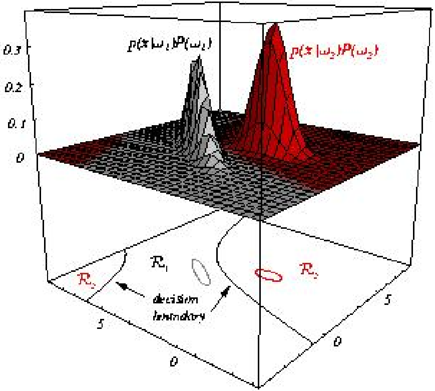

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Normal Density

- The structure of a Bayes classifier is determined primarily by the conditional densities $p(\vec{x}|\omega_i)$
- Usually assume normal distribution
- The uni-variate Normal density:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

for which

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} xp(x)dx = \mu \\ E[(x-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2 \end{aligned}$$

Normal Denity

- The uni-variate normal density is completely specified by 2 parameters, μ and σ.
- Multivariate Normal Density

$$p(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^t \Sigma^{-1} (\vec{x}-\vec{\mu})\right]$$

where $\vec{x}$ is a d-dim vector;
μ is a d-dim mean vector,
Σ is a $d \times d$ covariance matrix

- similarly,

$$\begin{aligned} \vec{\mu} &= E[\vec{x}] \\ \Sigma &= E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^t] \end{aligned}$$

Normal Density

- The covariance matrix Σ is always symmetric and positive semi-definite (i.e. Determinant $\geq$ 0)
- If $\Sigma = [\sigma_{ij}]$, the diagonal elements σ_{ii} is the variance of x_i, and the off-diagonal elements σ_{ij} the covariance of x_i, and x_j.
- If x_i and x_j are statistically independent, $\sigma_{ij} = 0$.
- The multi-variate normal density is completely specified by $d + d(d+1)/2$ parameters, d for the mean vectors, and $d(d+1)/2$ for the covariance matrix.

Some properties of Normal Densities

- Any linear combination of normally distributed random variables is also normal.
- Knowledge of the covariance matrix allows us to calculate the dispersion of the data in any direction – it determines the shape (hyperellipsoid) of the cluster of samples drawn from the distribution (fig. 2.9)

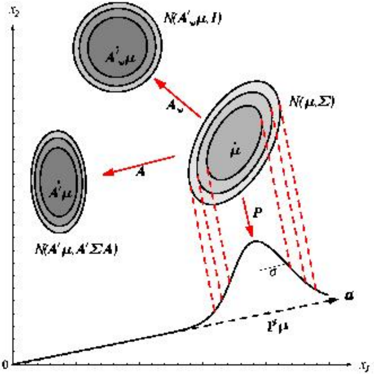

FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, **A**, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection **P** onto a line defined by vector **a**—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

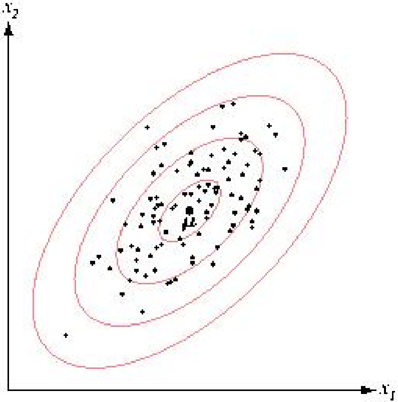

FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Properties of Normal Densities

- The principal axes of these hyperellipsoid are given by the eigenvectors of Σ.
- The Mahalanobis distance is $r = \left[(\vec{x} - \vec{\mu})^t \Sigma^{-1} (\vec{x} - \vec{\mu})\right]^{\frac{1}{2}}$
- The volume of the hyperellipsoid bounded by the equi-Mahalanobis contour r is

$$V = V_d |\Sigma|^{\frac{1}{2}} r^d$$

where V_d is the volume of a d-dim unit hypersphere

$$V_d = \begin{cases} \dfrac{\pi^{\frac{d}{2}}}{\left(\frac{d}{2}\right)!} & d \text{ even} \\ \dfrac{2^d \pi^{\frac{d-1}{2}} \left(\frac{d-1}{2}\right)!}{d!} & d \text{ odd} \end{cases}$$

Discriminant Function for Normal Density

- discriminant function for minimum-error-rate classification

$$g_i(\vec{x}) = \log p(\vec{x}|\omega_i) + \log P(\omega_i)$$

- Assume $p(\vec{x}|\omega_i) \sim N(\vec{\mu}_i, \Sigma_i)$, then

$$\begin{aligned} g_i(\vec{x}) &= -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| \\ &\quad + \log P(\omega_i) \\ g_i(\vec{x}) &= -\left[(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) + \log|\Sigma_i|\right] - 2\log P(\omega_i) \end{aligned}$$

  by dropping constant terms and multiply by +ve constant

- If the *a priori* probability for different classes the same, i.e. $P(\omega_i) = P(\omega_j)\ \forall i, j$, then

$$g_i(\vec{x}) = -\left[(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) + \log|\Sigma_i|\right]$$

Discriminant Function for Normal Density

There are 2 cases:

1. $\Sigma_i = \Sigma_j = \Sigma, \ \forall i, j$
2. Σ Arbitrary

Discriminant Function for Normal Density

case 1 $\Sigma_i = \Sigma_j = \Sigma, \quad \forall i, j$ then

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^t \Sigma^{-1} (\vec{x} - \vec{\mu}_i) + \log P(\omega_i), \qquad (1)$$

or

$$g_i(\vec{x}) = \frac{1}{2}(\vec{x} - \vec{\mu}_i)^t \Sigma^{-1} (\vec{x} - \vec{\mu}_i) \text{ if } P(\omega_i) = P(\omega_j) \forall i, j.$$

Expanding the quadratic form (eqn 1), we get

$$g_i(\vec{x}) = -\frac{1}{2}\left[\vec{x}^t \Sigma^{-1} \vec{x} - 2(\Sigma^{-1}\vec{\mu}_i)^t \vec{x} + \vec{\mu}_i^t \Sigma^{-1} \vec{\mu}_i\right] + \log P(\omega_i)$$

Note that $\vec{x}^t \Sigma^{-1} \vec{x}$ is a constant term, hence

$$g_i(\vec{x}) = (\Sigma^{-1}\vec{\mu}_i)^t \vec{x} + \left[-\frac{1}{2}\vec{\mu}_i^t \Sigma^{-1} \vec{\mu}_i + \log P(\omega_i)\right]$$

which is a linear discriminant function.
The decision boundary will be a hyperplane.

Discriminant Function for Normal Density

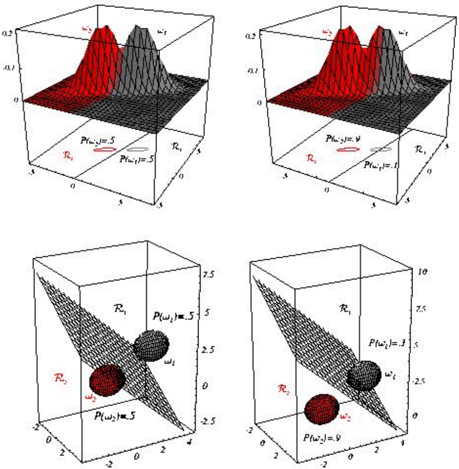

FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Discriminant Function for Normal Density

case 2 Arbitrary

$$g_i(\vec{x}) \;=\; -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

$$= \;\vec{x}^t\left[-\frac{1}{2}\Sigma_i\right]\vec{x} + (\Sigma_i^{-1}\vec{\mu}_i)^t\vec{x} - \frac{1}{2}\vec{\mu}_i^t\Sigma_i^{-1}\vec{\mu}_i - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

The decision surface are hyperquadrics, which can be

1. pair of hyperplanes
2. hypersphere
3. hyperellipsoid
4. hyperparaboloid
5. hyperhyperboloid

See fig 2.14 and fig 2.15.

Discriminant Function for Normal Density

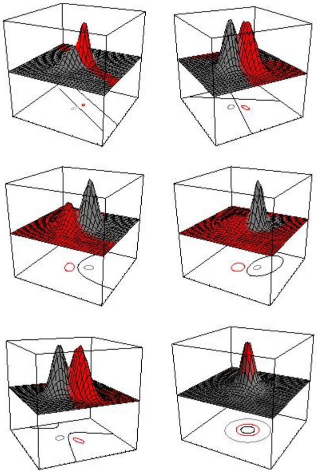

FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Discriminant Function for Normal Density

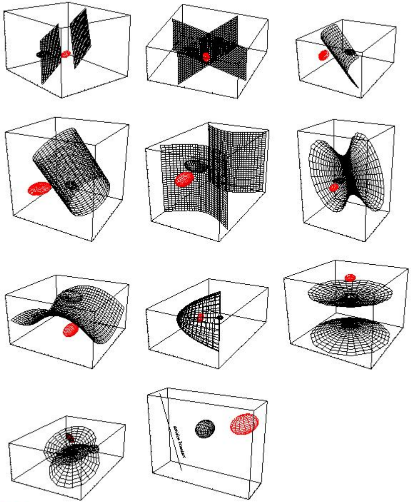

FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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