Bayesian risk classifier

1. Generalization

1. By allowing to use more than one feature – d dimensional feature space R_d
2. By allowing more than two states (categories)– $\omega_1, \ldots \omega_c$ states
3. By allowing more actions $\alpha_1 \ldots \alpha_a$ be the finite set of a possible actions.
4. By introducing loss functions

Bayesian risk classifier

Conditional Risk: Risk incurred in choosing decision α_i is

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_j|\omega_j) P(\omega_j|x) \tag{1}$$

The action for which risk $R(\alpha_i|x)$ is the minimum is chosen.

One-Zero loss function

Assume $\lambda(\alpha_i|\omega_j)$ written as λ_{ij}. So below we find the case as how to minimize the risk, i.e, choose the action α_i, for which $R(\alpha_i|x)$ is minimum

$$\lambda_{ij} = \begin{cases} 0 \text{ if } i == j \\ 1 \text{ otherwise} \end{cases} \tag{2}$$

$$R(\alpha_i|x) = \sum_{\forall j} \lambda(\alpha_{ij}) P(\omega_j|x) \tag{3}$$

$$= \sum_{i \neq j} P(\omega_j|x) \tag{4}$$

$$= 1 - P(\omega_i|x) - - - \text{ (one-zero loss function)} \tag{5}$$

The case when– for every correct decision, the loss function is 0 and for every incorrect decision, the error is 1, the Bayesian minimum Risk classifier is equal to Bayesian minimum Error classifier

The maximum criteria

The maximum criteria

1. Bayes error classifier: The class for which the posterior probability is maximum, is chosen
2. Bayes risk classifier: The action for which $-R(\alpha_i|x)$ is minimum

Discriminant functions

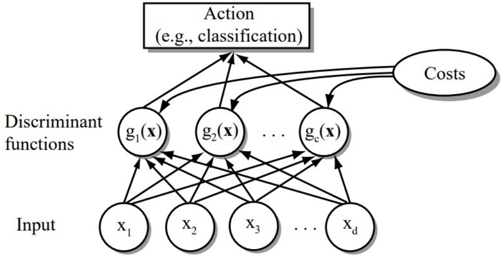


Figure: Functional structure of a general statistical pattern

Let $g_i(x)$, where $i = 1, 2, \ldots c$ be the set of discriminant functions. The classifier is said to assign a feature vector x to class ω_i if

$$g_i(x) \geq g_j(x), \text{for all } j \neq i \tag{6}$$

Discriminant functions

$$\text{Let, } g_i(x) = P(\omega_i|x) \tag{7}$$
$$= p(x|\omega_i)P(\omega_i) \tag{8}$$
$$\text{or, let } g_i(x) = f(P(\omega_i|x)) \tag{9}$$

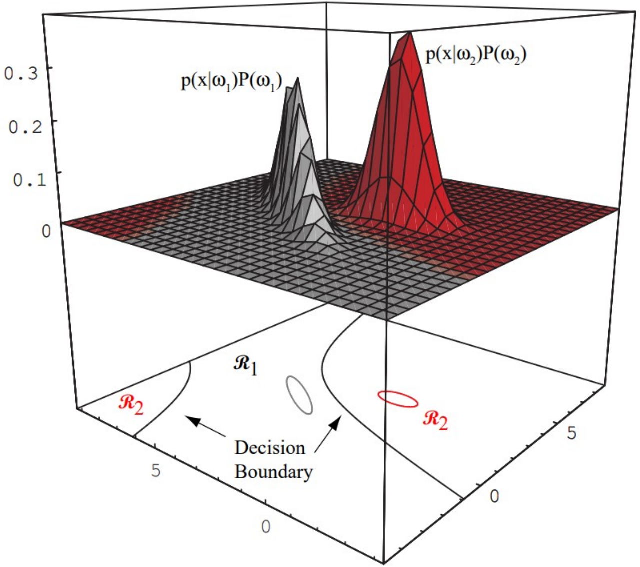
where $f(.)$ is a function of $P(\omega_i|x)$.

$$ln(P(\omega_i|x)) = ln(p(x|\omega_i)) + ln(P(\omega_i)) \tag{10}$$
$$g_i(x) = ln(p(x|\omega_i)) + ln(P(\omega_i)) \tag{11}$$

Decision region

1. The discriminant functions can be written in a variety of forms, but decision rules are equivalent
2. Assume, feature space divided into $\mathcal{R}_1, \mathcal{R}_2, \dots \mathcal{R}_c$ decision regions,
3. If $g_i(x) \geq g_j(x),$ for all $j \neq i$ $x \in \mathcal{R}_i$ and x will belong to class ω_i.

Decision region



The Normal Density

Now,lets see that if the probability density function is a Normal density function, then what will be the discriminant function and the decision boundary ?

- Univariate probability density
  Considering a feature component x,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{12}$$

The Normal Density

- Multivariate probability density Considering a d-dimensional feature vector x,

$$p(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \right] \tag{13}$$

$$\text{where, } \Sigma \text{ is the covariance matrix} \tag{14}$$

The class conditional probability is given as,

$$p(x|\omega_i) = \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i) \right] \tag{15}$$

The Normal Density

Based on Equation 11,

$$
\begin{aligned}
g_i(x) &= -\frac{d}{2} ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) + lnP(\omega_i) && (16)\\
&= -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) + lnP(\omega_i) && (17)
\end{aligned}
$$

The Normal Density

Let us examine the discriminant function and resulting classification for a number of special cases,

1. Case 1: When the features of class are statistically independent, i.e, $\Sigma_i = \sigma^2 I$
2. When the features of class may not be statistically independent but the covariance matrices for all of the classes are identical, i.e, $\Sigma_i = \Sigma$
3. the covariance matrices are different for each class, i.e, Σ_i is arbitrary

Case 1: $\Sigma_i = \sigma^2 I$

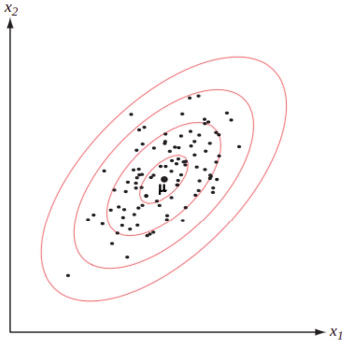


Figure: Functional structure of a general statistical pattern

The figure signify that the variance of different components of feature vector have same variance

Case 1: $\Sigma_i = \sigma^2 I$

Based on Equation 17,

$$g_i(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) + lnP(\omega_i) \tag{18}$$

$$= -\frac{1}{2\sigma^2}(x-\mu_i)^t(x-\mu_i) + lnP(\omega_i) \rightarrow \text{ Putting } \Sigma_i = \sigma^2 I \tag{19}$$

$$= -\frac{1}{2\sigma^2}[x^t x - 2x\mu_i^t + \mu_i^t\mu_i] + lnP(\omega_i) \tag{20}$$

$$= \frac{1}{\sigma^2}\mu_i^t x - \frac{1}{2\sigma^2}\mu_i^t\mu_i + lnP(\omega_i) \tag{21}$$

$$\approxeq w_i^t x + w_{i0} \rightarrow \text{Linear Expression} \tag{22}$$

where, $w_i = \frac{1}{\sigma^2}\mu_i$ and $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + lnP(\omega_i)$

Case 1: $\Sigma_i = \sigma^2 I$

On the decision boundary, $g_i(x) = g_j(x)$ So, the Equation for the decision boundary will be $g(x) = g_i(x) - g_j(x)$, Therefore $g(x) = 0$

$$g_i(x) = -\frac{1}{2\sigma^2}(x - \mu_i)^t(x - \mu_i) + lnP(\omega_i) \tag{23}$$

$$g_j(x) = -\frac{1}{2\sigma^2}(x - \mu_j)^t(x - \mu_j) + lnP(\omega_j) \tag{24}$$

Putting the above two Equations in $g(x) = g_i(x) - g_j(x)$, we obtain the equation for decision boundary $w^t(x - x_0) = 0$, where,

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j) \tag{25}$$

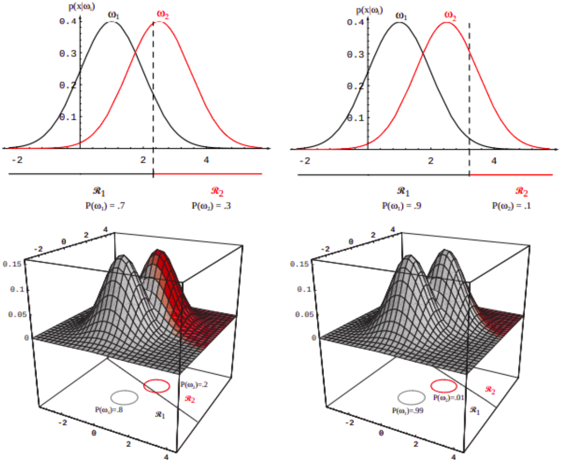
$$w = \mu_i - \mu_j \tag{26}$$

Case 1

The expression $w = \mu_i - \mu_j$, where μ_i is the mean of vectors taken from class ω_i and μ_j is the mean of vectors taken from class ω_j. Case 1 is summarized as

- In a 2-d, the decision boundary is linear,
- In a 3-d, the decision boundary is a plane,
- In multidimensional, the decision boundary is a hyperplane
- When $P(\omega_i) = P(\omega_j)$, $x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$ So, the decision surface is orthogonal to line joining μ_i and μ_j
- When $P(\omega_i) > P(\omega_j)$, decision surface will be shifted towards μ_j and vice versa.

Case 1



p(x|ω_i)
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
2
4
ℛ_1
ℛ_2
P(ω_1) = .7
P(ω_2) = .3
P(ω_1) = .9
P(ω_2) = .1
0.15
0.1
0.05
0
P(ω_2)=.2
P(ω_1)=.8
P(ω_2)=.01
P(ω_1)=.99

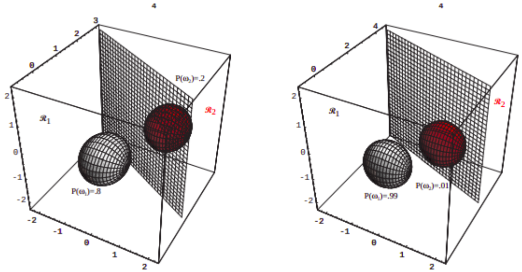


Figure 2.11: As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these 1-, 2- and 3-dimensional spherical Gaussian distributions.

Case 2: $\Sigma_i = \Sigma$

Because $\Sigma_i = \Sigma$, the Equation 16 becomes,

$$g_i(x) = -\frac{d}{2} ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2}(x - \mu_i)^t \Sigma^{-1}(x - \mu_i) + lnP(\omega_i) \tag{27}$$

The term $-\frac{d}{2} ln 2\pi - \frac{1}{2} \ln |\Sigma|$, is class independent, Therefore,

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \Sigma^{-1}(x - \mu_i) + lnP(\omega_i) \tag{28}$$

$$= -\frac{1}{2}[x^t \Sigma^{-1} x - 2\mu_i^t \Sigma^{-1} x + \mu_i^t \Sigma^{-1} \mu_i] + lnP(\omega_i) \tag{29}$$

$$= \mu_i^t \Sigma^{-1} x - \frac{1}{2}\mu_i^t \Sigma^{-1} \mu_i] + lnP(\omega_i) \tag{30}$$

Case 2

The above Equation is of the form

$$w_i^t x + w_{i0} \text{ where,} \tag{31}$$

$$w_i = \Sigma^{-1}\mu_i \tag{32}$$

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i] + lnP(\omega_i) \tag{33}$$

Case 2

The decision boundary will be

$$g(x) = g_i(x) - g_j(x) = 0 \tag{34}$$

where, $g_i(x)$ and $g_j(x)$ is given in Equation 30. Putting Equation 30, we get,

$$g(x) = (\mu_i^t \Sigma^{-1} - \mu_j^t \Sigma^{-1})x - \frac{1}{2}[\mu_i^t \Sigma^{-1} \mu_i - \mu_j^t \Sigma^{-1} \mu_j] + ln \frac{P(\omega_i)}{P(\omega_j)} = 0 \tag{35}$$

Simplifying the above equation,

$$g(x) = w^t (x - x_0) = 0 \tag{36}$$

$$w = \Sigma^{-1}(\mu_i - \mu_j) \tag{37}$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{ln \frac{P(\omega_i)}{p\omega_j}}{(\mu_i - \mu_j)^t \Sigma^{-1} (\mu_i - \mu_j)}(\mu_i - \mu_j) \tag{38}$$

Case 2

Case 2 is summarized as

- The discriminant function is linear
- when $P(\omega_i) = P(\omega_j)$,

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) \tag{39}$$

The decision surface is orthogonal but may not orthogonal to the line joining μ_i and μ_j.

Case 3: $\Sigma = arbitrary$

According to Equation 16,

$$g_i(x) = -\frac{d}{2} ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) + lnP(\omega_i) \quad (40)$$

$-\frac{d}{2} ln 2\pi$ is class independent but not $\frac{1}{2}\ln|\Sigma_i|$

$$g_i(x) \approx \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) + lnP(\omega_i) - \frac{d}{2} ln|\Sigma_i| \quad (41)$$

Therefore, $g_i(x) \approx x^t A_i x + B_i^t x + C_i$, where,

$$A_i = -\frac{1}{2}\Sigma_i^{-1} \quad (42)$$

$$B_i = \Sigma_i^{-1}\mu_i \quad (43)$$

$$C_i = -\frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) \quad (44)$$

Now, the discriminant function is quadratic

Case 3

$$g_i(x) \approx x^t A_i x + B_i^t x + C_i \tag{45}$$

$$g_j(x) \approx x^t A_j x + B_j^t x + C_j \tag{46}$$

Putting the above two into $g(x) = g_i(x) - g_j(x) = 0$, we obtain the decision boundary.

Case 3

Case 3 is summarized as

- the discriminant function is quadratic
- the decision boundary is quadratic

Bayes Decision Theory– Discrete

The feature vector x is discrete and binary.

$$x = \left[x_1, x_2, \ldots x_d \right] \tag{47}$$

where, $x_i = 0$ or $x_i = 1$

- Given two classes, $\omega_i = 0$ or 1,
- the different components are conditionally independent.
- p_i is the probability that the feature component $x_i = 1$, given that $\omega_i = 1$,
- $p_i = P_r[x_i = 1 | \omega_1]$ and $q_i = P_r[x_i = 1 | \omega_2]$
- If $p_i > q_i$ denote that x_i is more likely to have a value 1 if $x \in \omega_1$.

Bayes Decision Theory– Discrete

- So, the class conditional probability given the d-dimensional feature vector x, will be the product of all the independent probability values

$$p(x|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} \tag{48}$$

-

$$p(x|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i} \tag{49}$$