02_Classification_using_Bayes_Theory

2.1 Bayes Decision Theory 贝叶斯决策理论

ℹ Basic Assumptions

- The decision problem is posed in probabilistic terms.
- ALL relevant probability values are known.

2.1.1 Process

Given:

1. A test sample $\mathbf{x}$.
    - Contains features $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_l \end{bmatrix}$.
    - Often reduced, removed some non-discriminative (un-useful) features.
2. A list of classes/patterns $\omega = \{\omega_1, \omega_2, \dots \omega_c\}$.
    - Defined by human-being.
3. A classification method M.
   - A database storing multiple samples with the same type of x.
   - Each sample is assigned to an arbitrary class $\omega_{any} \in \{\omega_1, \omega_2, \dots \omega_c\}$.
   Do:

- $\{P(\omega_1|\mathbf{x}), \cdots, P(\omega_c|\mathbf{x})\} \leftarrow classify(M, \mathbf{x}, \omega)$
- That is, for all the possible classes, find:
  - The probability that the given x belongs to that class.
  Get:
- $\omega_{target}(\mathbf{x}) = \text{argmax}_i \left[P(\omega_i|x)\right], i \in [1, c]$.
- That is, assign x a class/pattern from ω with the most probable one.

Example
MNIST database.

- Test sample:
    - $x =$ A 28×28 grayscale image of a hand-written number.
- Set of classes:
    - ω = $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
- Classification Method:
    - Derived from 10,000 of 28×28 similar gray-scale images.

- Process:
  - Given an image, using the classification method, get a list of probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_c)\}$.
  - Select the ω_i with the largest probability $P(\omega_i)$, that is $selected = argmax[P(\omega_i)]$.

2.1.2 Properties of Variables.

- ℹ The set of all classes ω :
  - c available classes: $\omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$
- ℹ Prior Probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_c)\}$:
  - Probability Distribution of random variable ω_j in the database.
    - The fraction of samples in the database that belongs to class ω_j.
    - $P(\omega)$ is the prior knowledge on $\omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$.
  - It is Non-Negative.
    - $\forall i \in [1, c], P(\omega_i) \geq 0$.
    - The probabilities of all classes are greater-or-equal to 0.
  - It is Normalized.
    - $\sum_{i=1}^{c} P(\omega_i) = 1$.
    - The sum of the prior probabilities of all classes is 1.

2.2 Prior & Posterior Probabilities 先验与后验概率

2.2.1 Definition of Prior Probability 先验概率

- ℹ Decision *BEFORE* Observation (Naïve Decision Rule).
  - Don't care about test sample x.
  - Given x, always choose the class that:
    - has the most member in the database.
    - i.e., has the highest prior probability.
- Classification Process:
  1. $\omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$.
  2. By counting the number of members $Num(\omega_i)$ for each class $\omega_i \in \omega, i \in [1, c]$, we get the prior probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_c)\}$.
  3. Then, classify x directly into $\mathrm{argmax}_i[P(\omega_i)]$.
- The decision is the same all the time obviously, and the prob. of a right guess is $\frac{1}{c}$.

2.2.2 Definition of Posterior Probability 后验概率

- ℹ Decision *WITH* Observation.
  - Cares about test sample $\mathbf{x}$.

- Considering $\mathbf{x}$, as well as the prior probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \dots, P(\omega_c)\}$
  ,
    - and give $\mathbf{x}$ the class with the biggest posterior probability.

Posterior Probability of a class ω_j on test sample $\mathbf{x}$:
- Given test sample x,
- how possible does x could be classified into class ω_j.

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

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Pior}}{\text{Evidence}}$$

where:

Likelihood - $p(\mathbf{x}|\omega_j)$:
- *Known*
- The fraction of samples stored in the database that
    - is same to $\mathbf{x}$, and
    - is labeled to class ω_j.

Prior probability of class ω_j - $P(\omega_j)$:
- *Known*
- The fraction of samples stored in the database that
    - is not necessarily same to $\mathbf{x}$, and
    - is labeled to class ω_j.

Evidence - $P(\mathbf{x})$:
- *Irrelevant*
- Unconditional density of $\mathbf{x}$.
- $P(\mathbf{x}) = \sum_{j=1}^{N} p(\mathbf{x}|\omega_j) \cdot P(\omega_j)$

Special Cases

1. Equal Prior Probability

$$P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = \frac{1}{c}$$

- The amount of members in each class is same.
- Posterior probabilities $P(\omega_j|\mathbf{x})$ only depend on likelihoods $P(\mathbf{x}|\omega_j)$.

2. Equal Likelihood

$$P(\mathbf{x}|\omega_1) = P(\mathbf{x}|\omega_2) = \cdots = P(\mathbf{x}|\omega_c)$$

- The amount of members *that's same to $\mathbf{x}$* in each class is same.

- Posterior probabilities $P(\omega_j|\mathbf{x})$ only depend on priors $P(\omega_j)$.
- Back to Naïve Decision Rule.

2.2.3 Classification Examples

Given:

1. Test sample $x \in \{+, -\}$.
2. A list of classes $\omega = \{\omega_1 = cancer, \omega_2 = no\_cancer\}$.
3. Classification Method M, with known probabilities:
   - Prior Probabilities:
   - $P(\omega_1) = 0.008$
   - $P(\omega_2) = 1 - P(\omega_1) = 0.992$
   - Likelihoods:
   - For class $\omega_1 = cancer$: $P(+|\omega_1) = 0.98$, $P(-|\omega_1) = 0.02$
   - For class $\omega_2 = no\_cancer$: $P(+|\omega_2) = 0.03$, $P(-|\omega_2) = 0.97$.

   Classification:

- Given a test sample $x = +$.
  - The prob. that this person gets cancer is:
    - $P(\omega_1|+) = \dfrac{P(+|\omega_1) \times P(\omega_1)}{P(+)} = \dfrac{0.98 \times 0.008}{P(+)} = \dfrac{0.00784}{P(+)}$.
  - The prob. that this person doesn't gets cancer is:
    - $P(\omega_2|+) = \dfrac{P(+|\omega_2) \times P(\omega_2)}{P(+)} = \dfrac{0.03 \times 0.992}{P(+)} = \dfrac{0.02976}{P(+)}$
  - Therefore, the classification result would be:
    - $\omega_{target} = argmax_i[P(\omega_i|+)]$
      $= argmax_i[\dfrac{P(+|\omega_i) \times P(\omega_i)}{P(x)}]$
      $= argmax_i[P(+|\omega_i) \times P(\omega_i)]$
      $= \omega_2$, for $0.00784 < 0.02976$
    - That is, no_cancer.

2.3 Loss Functions 决策成本函数

2.3.0 Why do we use loss functions?

- Different selection errors may have differently significant consequences, i.e., "losses" or "costs". 不同决策的成本、后果不同。
  - In pure Naïve Bayes classification, we only consider probability.
  - However,
    - we can tolerate "non-cancer" being classified into "cancer",
    - while it's more lossy to classify "cancer" into "non-cancer".
  - There is a need to consider this kind of "loss" into our decision method.

- We want to know if the Bayes decision rule is optimal.
    - Need a evaluation method
    - calc how many error you make, sum together

2.3.1 Probability of Error

For only two classes:

- If $P(\omega_1|x) > P(\omega_2|x)$, $x \leftarrow \omega_1$. Prob. of error: $P(\omega_2|x)$.
- If $P(\omega_1|x) < P(\omega_2|x)$, $x \leftarrow \omega_2$. Prob. of error: $P(\omega_1|x)$.

2.3.2 Loss Function (i.e., "Cost Function")

Basics

- An action α_i for a given $\mathbf{x}$ is:
    - To assign the test pattern $\mathbf{x}$ with the class ω_i
- The loss $\lambda(\alpha_i|\omega_j)$ denotes the cost of:
    - Assigning a random test sample as ω_i,
    - while the actual class of the sample is ω_j.
    - For instance, $\lambda(\alpha_{\text{cancer}}|\omega_{\text{no\_cancer}})$ is the cost of diagnosing a patient without cancer as "having cancer".

Expected Loss & Bayes Risk

- Expected Loss (Average Loss, Conditional Risk) 期望成本
    - We don't actually know the true class of ω_j for a random sample $\mathbf{x}$, so we use the Expected Loss, i.e., the "average loss".
    - We consider the average loss of classifying a random sample into ω_i by considering:
        - For all class $\omega_j \in \omega$, the loss of classifying ω_i into ω_j, and
        - The probability that the random sample $\mathbf{x} \in \omega_j$, i.e., $P(\omega_j|\mathbf{x})$.

The expected loss of classifying a random sample $\mathbf{x}$ into ω_i is:

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

where:

- $\lambda(\alpha_i|\omega_j)$ is the cost of classifying $\mathbf{x}$ into ω_i with $\mathbf{x}$ belonging to ω_j actually.
- $P(\omega_i|\mathbf{x})$ is the (posterior) probability that $\mathbf{x}$ belongs to class ω_j.
    - Computed during Naïve Bayes Classification with $P(\omega_j)$ and $P(\mathbf{x}|\omega_j)$.
- Bayes Risk 贝叶斯风险
  - The modified measurement of the original Bayes Rule.

- Consider the importance of each error.
- Consider minimum loss, instead of maximum probability.
- Bayes Risk finds the action that gives the *minimum expected loss* classifying $\mathbf{x}$.

$$
\begin{aligned}
\alpha(\mathbf{x}) &= \operatorname{argmin}_{\alpha_i \in A} R(\alpha_i|\mathbf{x}) \\
&= \operatorname{argmin}_{\alpha \in A} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x})
\end{aligned}
$$

Derivation: A 2-class problem.

Given

- The test sample $\mathbf{x}$.
- Two classes: $\omega = \{\omega_1, \omega_2\}$
- Calculated posterior probabilities during Naive Bayes:
  - $P(\omega_1|\mathbf{x})$, $P(\omega_2|\mathbf{x})$
- Loss Matrix:
  - $\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$
  - where $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$

Do

- $\omega^* = \operatorname{argmin}_{\alpha_i \in A} R(\alpha_i|\mathbf{x})$
- The condition of choosing α_1 is:

$$
\begin{aligned}
& R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x}) \\
\iff & \lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x}) < \lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x}) \\
\iff & (\lambda_{21} - \lambda_{11}) \cdot P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22}) \cdot P(\omega_2|\mathbf{x}) \\
\iff & \frac{P(\omega_1|\mathbf{x})}{P(\omega_2|\mathbf{x})} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \\
\iff & \frac{P(\mathbf{x}|\omega_1) \cdot P(\omega_1)}{P(\mathbf{x}|\omega_2) \cdot P(\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \\
\iff & \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22}) \cdot P(\omega_2)}{(\lambda_{21} - \lambda_{11}) \cdot P(\omega_1)} = \theta
\end{aligned}
$$

2.3.3 Examples

Minimum Prob. Error and Minimum Risk

Remark: The Gaussian Distribution.

$$
x \in \mathbb{R} \sim Gaussian(\mu, \sigma): \ P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
$$

Given:

- Random distributions of samples in 2 classes ω_1 and ω_2 respectively.
  - ω_1: $\mu = 0$, $\sigma = \frac{1}{\sqrt{2}} \implies P(x|\omega_1) = \frac{1}{\sqrt{\pi}} e^{-x^2}$
  - ω_2: $\mu = 1$, $\sigma = \frac{1}{\sqrt{2}} \implies P(x|\omega_2) = \frac{1}{\sqrt{\pi}} e^{-(x-1)^2}$
- Loss Matrix:
  - $\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1.0 \\ 0.5 & 0 \end{pmatrix}$

  Do:

Minimum Error

The threshold value x_0 where the two distributions are equal

- i.e., minimum probability of error

$$P(x_0|\omega_1) = P(x_0|\omega_2)$$

$$\implies \frac{1}{\sqrt{\pi}} e^{-x_0^2} = \frac{1}{\sqrt{\pi}} e^{-(x_0-1)^2}$$

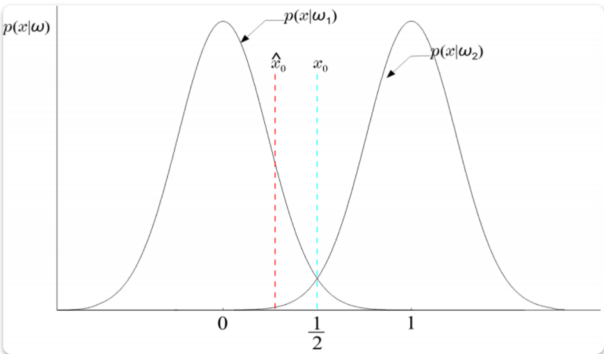
$$\implies x_0 = -x_0 + 1$$

$$\implies x_0 = \frac{1}{2}$$

Minimum Risk

The threshold $\hat{x}_0$ for minimum $R(\alpha_i|x)$.

$$
\begin{aligned}
& R(\alpha_1|\hat{x}_0) = R(\alpha_2|\hat{x}_0) \\
\Longrightarrow\ & \lambda_{11}\cdot P(\omega_1|\hat{x}_0) + \lambda_{12}\cdot P(\omega_2|\hat{x}_0) = \lambda_{21}\cdot P(\omega_1|\hat{x}_0) + \lambda_{22}\cdot P(\omega_2|\hat{x}_0) \\
\Longrightarrow\ & (\lambda_{21}-\lambda_{11})\cdot P(\omega_1|\hat{x}_0) = (\lambda_{12}-\lambda_{22})\cdot P(\omega_2|\hat{x}_0) \\
\Longrightarrow\ & \frac{P(\omega_1|\hat{x}_0)}{P(\omega_2|\hat{x}_0)} = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \\
\Longrightarrow\ & \frac{P(\hat{x}_0|\omega_1)\cdot P(\omega_1)}{P(\hat{x}_0|\omega_2)\cdot P(\omega_2)} = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \\
\Longrightarrow\ & \frac{P(\hat{x}_0|\omega_1)}{P(\hat{x}_0|\omega_2)} = \frac{(\lambda_{12}-\lambda_{22})\cdot P(\omega_2)}{(\lambda_{21}-\lambda_{11})\cdot P(\omega_1)} \\
\Longrightarrow\ & \frac{P(\hat{x}_0|\omega_1)}{P(\hat{x}_0|\omega_2)} = \frac{(1-0)\times\frac{1}{2}}{(0.5-0)\times\frac{1}{2}} = 2 \\
\Longrightarrow\ & P(\hat{x}_0|\omega_1) = 2P(\hat{x}_0|\omega_2) \\
\Longrightarrow\ & \frac{1}{\sqrt{\pi}}e^{-\hat{x}_0^2} = \frac{2}{\sqrt{\pi}}e^{-(\hat{x}_0-1)^2} \\
\Longrightarrow\ & -\hat{x}_0^2 = \ln 2 - \hat{x}_0^2 + 2\hat{x}_0 - 1 \\
\Longrightarrow\ & 2\hat{x}_0 = 1 - \ln 2 \\
\Longrightarrow\ & \hat{x}_0 = \frac{1-\ln 2}{2}
\end{aligned}
$$



2.4 Discriminant Functions 判别函数

2.4.1 Definition of Discriminant Function

A Discriminant Function is a function f that satisfies the following property:

- If:
  - $f(\cdot)$ monotonically increases, and
  - $\forall i \neq j,\ f\Big(P(\omega_i|\mathbf{x})\Big) > f\Big(P(\omega_j|\mathbf{x})\Big)$

- Then:
    - $\mathbf{x} \leftarrow \omega_i$
- That is, the function is able to "tell", or "discriminate" a certain ω_i from others.
    - i.e., it separates ω_i and $\neg\omega_i$.

A sample usage of a discriminant function: Given two classes ω_i and ω_j, define $g(\mathbf{x}) \equiv P(\omega_i|\mathbf{x}) - P(\omega_j|\mathbf{x}) = 0$.

- $g(\mathbf{x}) = 0$: Decision Surface;
- $g(\mathbf{x}) > 0$: Region R_i where $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$;
- $g(\mathbf{x}) < 0$: Region R_i where $P(\omega_i|\mathbf{x}) < P(\omega_j|\mathbf{x})$;

2.4.2 Property of Discriminant Function

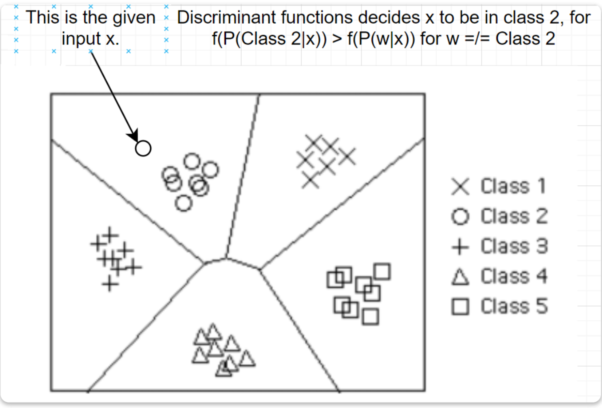
1. One function per class.
    1. A discriminant function is able to "tell" a certain one ω_i specifically for any input x.
2. Various discriminant functions $\rightarrow$ Identical classification results. 样式各异，结果相同。
    1. It is correct to say, the discriminant functions:
        1. *Preserves* the original monotonically increase of its inputs.
        2. But changes the changing rate by *processing* the inputs.
    2. i.e.,
        1. "$\forall i \neq j, f(g_i(x)) > f(g_j(x)) \wedge f \nearrow$ "and "$\forall i \neq j, g_i(x) > g_j(x)$" are equivalent in decision.
        2. Changing growth rate of input:
            1. $f(g_i(x)) = k \cdot g_i(x)$, a linear change.
            2. $f(g_i(x)) = \ln g_i(x)$, a log change, i.e., it grows, but slower as it proceed.
        3. Therefore, the discriminant function may vary, but the output is always the same.
3. Examples of discriminant functions:
    1. Minimum Risk: $g_i(x) = -R(\alpha_i|x) = -\lambda(\alpha_i|x) \times P(\omega_i|x)$, for $i \in [1, c]$
    2. Minimum Error Rate: $g_i(x) = P(\omega_i|x)$, for $i \in [1, c]$

2.4.3 Decision Region 决策区域

- c discriminant functions $\implies$ c decision regions
    - $g_i(x) \implies R_i \subset R^d, i \in [1, c]$
- One function per decision region that is distinct and mutual-exclusive.
    - A decision region is defined as: $R_i = \{x|x \in R^d : \forall i \neq j, g_i(x) > g_j(x)\}$, where
    - $\forall i \neq j, R_i \cap R_j = \emptyset$, and $\cup_{i=1}^{c} R_i = R^d$

2.4.4 Decision Boundaries 决策边界

- "Surface" in feature space, where ties occur among 2 or more largest discriminant functions.
- x_0 is on the decision boundary/surface if and only if
  - $\exists \omega_i, \omega_j \in \omega, g_i(x_0) = g_j(x_0)$.



2.5 Bayesian Classification for Normal Distributions

2.5.1 Multi-Dimensional Normal Distribution 高维正态分布

1-D Case 多类别，一维数据

- There are several classes:
  - Each class has its own distribution of data samples.
  - i.e., each class has its own μ and σ.
- For a specific class, there are plenty of data samples:
  - Each sample is a *scalar*, that is a 1×1 "matrix", which is a "plain number".
  - The samples follows a Normal Distribution.

Suppose data samples in a specific class ω_i conforms a normal distribution:

$$x \sim N(\mu_i, \sigma_i) : P(x|\omega_i) = \frac{1}{\sigma_i \sqrt{2\pi}}^{\frac{(x-\mu_i)^2}{2\sigma^2}}$$

where:

- μ_i is the mean value, $\mu_i = E(x)$
- σ_i^2 is the variance, $\sigma_i^2 = E\left[(x-\mu)^2\right]$

Multivariate Case 多类别，高维数据

- There are several classes:
  - Each class has its own distribution of data samples,
  - i.e., each class has its own μ and σ.
- For a specific class, there are plenty of data samples:
  - Each sample is a *vector*, that is a $d \times 1$ matrix, where d is the dimension of data.

- The samples follow a d-dimensional Normal Distribution.

Suppose data samples in a specific class ω_i conforms a normal distribution:

$$\mathbf{x} \sim N(\mu_i, \Sigma_i): \ P(\mathbf{x}|\omega_i) = \frac{1}{|\Sigma_i|^{\frac{1}{2}} \cdot (2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)}$$

Regular Variables:

- d-dimensional random variable: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$
- d-dimensional mean vector: $\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{id} \end{bmatrix} = \begin{bmatrix} E(x_{i1}) \\ E(x_{i2}) \\ \vdots \\ E(x_{id}) \end{bmatrix}$
- $d \times d$ covariate matrix: $\sigma_i = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1d} \\ \sigma_{i21} & \sigma_{i22} & \cdots & \sigma_{i2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{id1} & \sigma_{id2} & \cdots & \sigma_{idd} \end{pmatrix} = E\Big[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^\top\Big]$

Explanation of exponent $-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)$:

- $(X-\mu_i)^\top = [(x_1-\mu_{i1}) \quad (x_2-\mu_{i2}) \quad \cdots \quad (x_d-\mu_{id})]$
- $\Sigma_i^{-1} = \begin{pmatrix} \sigma'_{i11} & \sigma'_{i12} & \cdots & \sigma'_{i1d} \\ \sigma'_{i21} & \sigma'_{i22} & \cdots & \sigma'_{i2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma'_{id1} & \sigma'_{id2} & \cdots & \sigma'_{idd} \end{pmatrix}$, the inverse of the covariance matrix.
- $(X-\mu_i) = \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ \cdots \\ x_d-\mu_{id} \end{bmatrix}$

The exponent as a whole:

$$
-\frac{1}{2}(X-\mu_i)^\top\Sigma_i^{-1}(X-\mu_i)
$$

$$
=-\frac{1}{2}[(x_1-\mu_{i1}) \quad (x_2-\mu_{i2}) \quad \cdots \quad (x_d-\mu_{id})]\begin{pmatrix}\sigma'_{i11} & \sigma'_{i12} & \cdots & \sigma'_{i1d}\\ \sigma'_{i21} & \sigma'_{i22} & \cdots & \sigma'_{i2d}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma'_{id1} & \sigma'_{id2} & \cdots & \sigma'_{idd}\end{pmatrix}\begin{bmatrix}x_1-\mu_{i1}\\ x_2-\mu_{i2}\\ \cdots\\ x_d-\mu_{id}\end{bmatrix}
$$

$$
=-\frac{1}{2}[a_1 \quad a_2 \quad \cdots a_d]\begin{bmatrix}x_1-\mu_{i1}\\ x_2-\mu_{i2}\\ \cdots\\ x_d-\mu_{id}\end{bmatrix}
$$

$$
=y\geq 0
$$

Example: 2-D Case

$$
\mathbf{x}\sim N(\mu_i,\sigma_i):P(\mathbf{x}|\omega_i=\frac{1}{|\Sigma_i|^{\frac{1}{2}}\cdot(2\pi)}e^{-\frac{1}{2}(x_1-\mu_{i1} \quad x_2-\mu_{i2})\Sigma_i^{-1}\begin{pmatrix}x_1-\mu_{i1}\\ x_2-\mu_{i2}\end{pmatrix}}
$$

where:

- 2-dimensional random variable: $\mathbf{x}=\begin{pmatrix}x_1\\ x_2\end{pmatrix}$.
- 2-dimensional mean vector: $\mu_i=\begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}$
- 2×2 covariate matrix Σ_i:

$$
\Sigma_i=E\Big[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^\top\Big]
$$

$$
=E\Big[\begin{pmatrix}x_1-\mu_{i1}\\ x_2-\mu_{i2}\end{pmatrix}(x_1-\mu_{i1} \quad x_2-\mu_{i2})\Big]
$$

$$
=E(\begin{bmatrix}(x_1-\mu_{i1})^2 & (x_1-\mu_{i1})(x_2-\mu_{i2})\\ (x_1-\mu_{i1})(x_2-\mu_{i2}) & (x_2-\mu_{i2})^2\end{bmatrix})
$$

$$
=\begin{bmatrix}E\Big[(x_1-\mu_{i1})^2\Big] & E\Big[(x_1-\mu_{i1})(x_2-\mu_{i2})\Big]\\ E\Big[(x_1-\mu_{i1})(x_2-\mu_{i2})\Big] & E\Big[(x_2-\mu_{i2})^2\Big]\end{bmatrix}
$$

$$
=\begin{bmatrix}\sigma_1^2 & \sigma\\ \sigma & \sigma_2^2\end{bmatrix}
$$

2.5.2 Minimum-error-rate classification

Recall:

- Minimum-error-rate means that we ignore the "cost" of each decision.
- In other words, we only select the classes based on probabilities.

Pattern of Discriminant Function

The discriminant function of MER classification could be given by:

$$\forall i \in [1, c] \cap \mathbb{N}^+,\ g_i(\mathbf{x}) = \ln P(\omega_i|\mathbf{x})$$

Namely,

$$\begin{aligned}
& g_i(\mathbf{x}) = \ln P(\omega_i|\mathbf{x}) \\
\Longrightarrow\ & g_i(\mathbf{x}) = \ln\Big[P(\mathbf{x}|\omega_\mathbf{i}) \cdot P(\omega_i)\Big] \\
\Longrightarrow\ & g_i(\mathbf{x}) = \ln\Big[P(\mathbf{x}|\omega_i)\Big] + \ln\Big[P(\omega_i)\Big] \\
\Longrightarrow\ & g_i(\mathbf{x}) = \ln\Big[\frac{1}{|\Sigma_i|^{\frac{1}{2}} \cdot (2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)}\Big] + \ln\Big[P(\omega_i)\Big] \\
\Longrightarrow\ & g_i(\mathbf{x}) = \ln\Big[\frac{1}{|\Sigma_i|^{\frac{1}{2}} \cdot (2\pi)^{\frac{d}{2}}}\Big] - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big] \\
\Longrightarrow\ & g_i(\mathbf{x}) = \Big(-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i|\Big) - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big]
\end{aligned}$$

Here, $-\frac{d}{2}\ln(2\pi)$ is a constant, which could be ignored.

★ The discriminant function is then updated as:

$$g_i(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big]$$

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

That is:

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{|\omega|} = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

- All the classes have a *Common Covariance Matrix* of $\sigma^2 I$.
- The common covariate matrix is *isotropic (各向同性的)* with respect to any class.
    - i.e., the variance is the same in all directions.
    - i.e., no directional preference in the spread of distribution

Therefore, we have:

$$|\Sigma_i| = \sigma^{2d}$$

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

This is the original discriminant function:

$$g_i(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big]$$

Here, $-\frac{1}{2}\ln|\Sigma_i| = -\frac{1}{2}\ln|\sigma^2 I|$ is a constant, therefore can be ignored:

$$\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}(\mathbf{x}-\mu_i)^\top \cdot (\frac{1}{\sigma^2}I) \cdot (\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{(\mathbf{x}-\mu_i)^\top(\mathbf{x}-\mu_i)}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{(\mathbf{x}^\top-\mu_i^\top)(\mathbf{x}-\mu_i)}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{\mathbf{x}^\top\mathbf{x} - \mathbf{x}^\top\mu_i - \mu_i^\top\mathbf{x} + \mu_i^\top\mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{\mathbf{x}^\top\mathbf{x} - 2\mu_i^\top\mathbf{x} + \mu_i^\top\mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{\|\mathbf{x}-\mu_i\|^2}{2\sigma^2} + \ln\Big[P(\omega_i)\Big]
\end{aligned}$$

Note: $\|\cdot\|^2$ denotes the *Euclidean Distance.*
Moreover $\mathbf{x}^\top\mathbf{x}$ is the same across all classes, therefore can be ignored:

$$\begin{aligned}
g_i(\mathbf{x}) &= -\frac{-2\mu_i^\top\mathbf{x} + \mu_i^\top\mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= \frac{\mu_i^\top\mathbf{x}}{\sigma^2} - \frac{\mu_i^\top\mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] \\
&= \Big(\frac{\mu_i}{\sigma^2}\Big)^\top\mathbf{x} + \Big(-\frac{\mu_i^\top\mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big]\Big)
\end{aligned}$$

Namely,

$$g_i(\mathbf{x}) = \mathbf{w}_i^\top\mathbf{x} + w_{i0}$$

where:

- $\mathbf{w}_i = \frac{\mu_i}{\sigma^2}$ is the weight vector, and

- $w_{i0} = -\frac{\mu_i^\top \mu_i}{2\sigma^2} + \ln\Big[P(\omega_i)\Big]$ is the threshold / bias scalar.

This is a Linear Discriminant Function. The decision surface is thus:

$$
\begin{aligned}
& g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0 \\
\implies & \mathbf{w}_i^\top \mathbf{x} + w_{i0} - (\mathbf{w}_j^\top \mathbf{x} + w_{j0}) = 0 \\
\implies & (\mathbf{w}_i - \mathbf{w}_j)^\top \mathbf{x} + (w_{i0} - w_{j0}) = 0 \\
\implies & (\frac{\mu_i - \mu_j}{\sigma^2})^\top \mathbf{x} + (w_{i0} - w_{j0}) = 0 \\
\implies & (\mu_i - \mu_j)^\top \mathbf{x} + \sigma^2 (w_{i0} - w_{j0}) = 0 \\
\implies & (\mu_i - \mu_j)^\top \mathbf{x} + \sigma^2 (\frac{-\mu_i^\top \mu_i}{2\sigma^2} - \frac{-\mu_j^\top \mu_j}{2\sigma^2} + \ln\Big[P(\omega_i)\Big] - \ln\Big[P(\omega_j)\Big]) = 0 \\
\implies & (\mu_i - \mu_j)^\top \mathbf{x} - \frac{1}{2}(\mu_i^\top \mu_i - \mu_j^\top \mu_j) + \sigma^2 \ln\Big[\frac{P(\omega_i)}{P(\omega_j)}\Big] = 0 \\
\implies & (\mu_i - \mu_j)^\top \mathbf{x} - \frac{1}{2}(\mu_i - \mu_j)^\top (\mu_i + \mu_j) + \sigma^2 \ln\Big[\frac{P(\omega_i)}{P(\omega_j)}\Big] = 0 \\
\implies & \mathbf{x} - \frac{1}{2}(\mu_i + \mu_j) + \sigma^2 \ln\Big[\frac{P(\omega_i)}{P(\omega_j)}\Big] \cdot \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2} = 0 \\
\implies & \mathbf{x} = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 \ln\Big[\frac{P(\omega_i)}{P(\omega_j)}\Big] \cdot \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2} \in \mathbb{R}^2
\end{aligned}
$$

Case II: $\Sigma_i = \Sigma$

That is:

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{|\omega|} = \Sigma$$

- All the classes have a *Common Covariance Matrix* of Σ.
- More general than Case I.

This is the original discriminant function:

$$g_i(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \mu_i)^\top \Sigma_i^{-1}(\mathbf{x} - \mu_i) + \ln\Big[P(\omega_i)\Big]$$

Here, $-\frac{1}{2}\ln|\Sigma_i| = -\frac{1}{2}\ln|\Sigma|$ is constant, which could be ignored:

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}(\mathbf{x}^\top - \mu_i^\top)\Big(\Sigma^{-1}\mathbf{x} - \Sigma^{-1}\mu_i\Big) + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}(\mathbf{x}^\top \Sigma^{-1}\mathbf{x} - \mathbf{x}^\top \Sigma^{-1}\mu_i - \mu_i^\top \Sigma^{-1}\mathbf{x} + \mu_i^\top \Sigma^{-1}\mu_i) + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}(\mathbf{x}^\top \Sigma^{-1}\mathbf{x} - 2\mu_i^\top \Sigma^{-1}\mathbf{x} + \mu_i^\top \Sigma^{-1}\mu_i) + \ln\Big[P(\omega_i)\Big]
\end{aligned}
$$

Here, $\mathbf{x}^\top \Sigma^{-1}\mathbf{x}$ is the same across all classes, thus can be ignored:

$$
g_i(\mathbf{x}) = \mu_i^\top \Sigma^{-1}\mathbf{x} + (-\frac{\mu_i^\top \Sigma^{-1}\mu_i}{2} + \ln\Big[P(\omega_i)\Big])
$$

Namely,

$$
g_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x} + w_{i0}
$$

where:

- $\mathbf{w}_i = \mu_i$ is the weight vector;
- $w_{i0} = -\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \ln\Big[P(\omega_i)\Big]$ is the threshold / bias scalar.

Case III: Σ_i is arbitrary

In most cases, for each class ω_i, Σ_i, the covariance/spread of data in this class is arbitrary. This is the original discriminant function:

$$
g_i(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln\Big[P(\omega_i)\Big]
$$

We can derive that:

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu_i)^\top\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}(\mathbf{x}^\top - \mu_i^\top)(\Sigma_i^{-1}\mathbf{x} - \Sigma_i^{-1}\mu_i) + (-\frac{1}{2}|\Sigma_i| + \ln\Big[P(\omega_i)\Big]) \\
&= -\frac{1}{2}(\mathbf{x}^\top\Sigma_i^{-1}\mathbf{x} - \mathbf{x}^\top\Sigma_i^{-1}\mu_i - \mu_i^\top\Sigma_i^{-1}\mathbf{x} + \mu_i^\top\Sigma_i^{-1}\mu_i) + (-\frac{1}{2}|\Sigma_i| + \ln\Big[P(\omega_i)\Big]) \\
&= -\frac{1}{2}(\mathbf{x}^\top\Sigma_i^{-1}\mathbf{x} - 2\mu_i^\top\Sigma_i^{-1}\mathbf{x} + \mu_i^\top\Sigma_i^{-1}\mu_i) + (-\frac{1}{2}|\Sigma_i| + \ln\Big[P(\omega_i)\Big]) \\
&= -\frac{1}{2}\mathbf{x}^\top\Sigma_I^{-1}\mathbf{x} + \mu_i^\top\Sigma_i^{-1}\mathbf{x} - \frac{1}{2}\mu_i^\top\Sigma_i^{-1}\mu_i - \frac{1}{2}|\Sigma_i| + \ln\Big[P(\omega_i)\Big] \\
&= -\frac{1}{2}\mathbf{x}^\top\Sigma_i^{-1}\mathbf{x} + \mu_i^\top\Sigma_i^{-1}\mathbf{x} + (-\frac{\mu_i^\top\Sigma_i^{-1}\mu_i + |\Sigma_i|}{2} + \ln\Big[P(\omega_i)\Big]) \\
&= \mathbf{x}^\top(-\frac{1}{2}\Sigma_i^{-1})\mathbf{x} + (\mu_i^\top\Sigma_i^{-1})\mathbf{x} + (-\frac{\mu_i^\top\Sigma_i^{-1}\mu_i + |\Sigma_i|}{2} + \ln\Big[P(\omega_i)\Big])
\end{aligned}
$$

Namely,

$$
g_i(\mathbf{x}) = \mathbf{x}^\top\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^\top\mathbf{x} + w_{i0}
$$

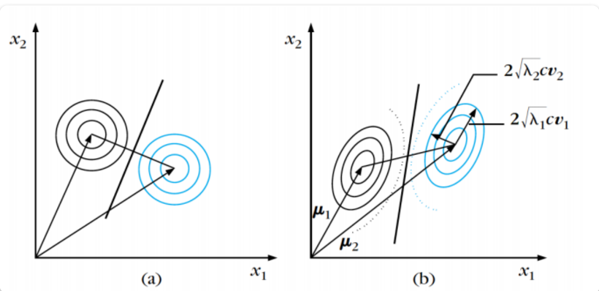
where:

- $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$ is the Quadratic matrix.
- $\mathbf{w}_i = \mu_i^\top\Sigma_i^{-1}$ is the weight vector.
- $w_{i0} = -\frac{\mu_i^\top\Sigma_i^{-1}\mu_i + |\Sigma_i|}{2} + \ln\Big[P(\omega_i)\Big]$ is the threshold / bias scalar.

Summary

Again, for special covariance matrices:

- $\Sigma_i = \sigma^2 I$:
  - Assign x to ω_i if there is a smaller Euclidean Distance: $d_{Euclidean} = \|X - \mu_i\|$
- $\Sigma_i = \Sigma$:
  - Assign x to ω_i if there is a smaller Mahalanobis Distance: $d_{Mahalanobis} = \sqrt{(X-\mu_i)^\top\Sigma^{-1}(X-\mu_i)}$



2.5.3 Examples

Given:

- Two classes: ω_1, ω_2
- Prior probabilities:
  - $P(\omega_1) = P(\omega_2)$.
- Posterior probabilities:

  - $P(\mathbf{x}|\omega_1) \sim N(\mu_1, \Sigma)$
  - $P(\mathbf{x}|\omega_2) \sim N(\mu_2, \Sigma)$
  - where:
  - $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$
  - $\Sigma = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$

  Do:

- Classify $\mathbf{x} = \begin{pmatrix} 1.0 \\ 2.2 \end{pmatrix}$ using Bayes Classification.

  Solve:

  Compute inverse of covariance matrix:

$$\begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}^{-1} = \frac{1}{1.1 \times 1.9 - 0.3^2} \begin{pmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{pmatrix}$$

$$= \frac{1}{2} \begin{pmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{pmatrix}$$

$$\Sigma^{-1} = \begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix}$$

Compute *Mahalanobis distance* using μ_1 and μ_2.

$$d^2(\mathbf{x}, \mu_i) = (\mathbf{x} - \mu_i)^\top \Sigma^{-1} (\mathbf{x} - \mu_i)$$

$$\mathbf{x} - \mu_1 = \begin{pmatrix} 1.0 \\ 2.2 \end{pmatrix}, \; \mathbf{x} - \mu_2 = \begin{pmatrix} -2.0 \\ -0.8 \end{pmatrix}$$

$$d^2(\mathbf{x}, \mu_1) = (1.0 \quad 2.2) \begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix} \begin{pmatrix} 1.0 \\ 2.2 \end{pmatrix} = 2.952$$

$$d^2(\mathbf{x}, \mu_2) = (-2.0 \quad -0.8) \begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix} \begin{pmatrix} -2.0 \\ -0.8 \end{pmatrix} = 3.672$$

Therefore, classify $\mathbf{x} \leftarrow \omega_1$, since $d^2(\mathbf{x}, \mu_2) > d^2(\mathbf{x}, \mu_1)$.