CHAPTER 3

1 # PATTERN RECOGNITION AND BAYES OPTIMAL CLASSIFIER

- **Statistical Pattern Recognition** - is concerned with classification of objects on the basis of quantitative features.

→ patterns

- **Mathematical Description of Patterns**

→ features :

→ feature space :

→ feature vectors:

→ realization :

→ decision :

- **Feature ~~Extraction~~ Extraction and Feature selection**

→ feature extraction

→ feature selection

- **Example of Classification Problem**

→ Camera shot (original image)

→ Preprocessing of the image (noise filtering, calculation of greyscale

→ feature extraction and selection

→ Classification

- **Statistical Model of Classification**

→ An object belongs to one of the K classes represented by the hypotheses $H_1, \dots, H_K$ which are characterized by prior probabilities $\pi_1, \dots, \pi_K$

→ true class label cannot be observed directly, instead a feature vector x is measured which is interpreted as a realization of the ~~rand~~ random vector X.

→ For an object from class k, the random vector X is distributed according to the class conditional pdf f_k.

→ A classifier $\delta : \mathbb{R}^D \to \{1, \dots, K\}$, $x \mapsto \delta(x)$ estimates true class on basis of x.

→ Goal: construct an optimal classifier δ^* which minimizes the probability of error.

- **Classifier and Decision regions**

→ A classifier $\delta : \mathbb{R}^D \to \{1, \dots, K\}$, $x \mapsto \delta(x)$

corresponds to a partition of feature space $\mathbb{R}^D$ into mutually exclusive subsets or decision regions $R_k \subseteq \mathbb{R}^D$ where

$$\bigcup_{k=1}^{K} R_k = \mathbb{R}^D, \quad R_i \cap R_j \neq 0 \text{ for } i \neq j$$

and

$$\delta(x) = \begin{cases} 1 & \text{if } x \in R_1 \\ 2 & \text{if } x \in R_2 \\ \vdots & \vdots \\ K & \text{if } x \in R_K \end{cases}$$

→ error free classification is only possible if the sets

$$S_k = \{ x \in \mathbb{R}^D \mid f_k(x) > 0 \} \qquad k = 1 \dots K$$

are mutually exclusive. i.e class conditional densities $f_k(x)$ do not overlap.

→ In this case, an error free classifier is obtained via $R_K = S_K$ i.e

$$\delta(x) = k \iff x \in S_K$$

- **Summary of different probability distributions**

1. Prior probabilities of the classes : $\pi_1, \dots, \pi_K$
2. Class conditional densities of feature vectors $f_k(x)$
3. Joint distribution over pairs $f(x,k) = \pi_k \cdot f_k(x)$

- **Derived Probability Distributions**

→ From joint distribution $f(x,k)$, the marginal distribution of the feature vectors (independent of the class) is obtained as :

$$\boxed{f(x) = \sum_{k=1}^{K} f(x,k) = \sum_{k=1}^{K} \pi_k \cdot f_k(x)}$$

→ Accordingly, the posterior distribution of classes is given by conditional probabilities

$$\boxed{p(k|x) = \frac{f(x,k)}{f(x)} = \frac{\pi_k \cdot f_k(x)}{\sum_{j=1}^{k} \pi_j \cdot f_j(x)}}$$

→ The posterior distribution is normalized i.e

$$\boxed{\sum_{k=1}^{K} p(k|x) = 1}$$

- **Bayes Optimal Classifier**

→ The classifier δ^* which minimizes the probability of error is called Bayes optimal classifier which is obtained as:

$$\boxed{\delta^*(x) = \underset{k=1,\dots,K}{\operatorname{argmax}}\ p(k|x) = \underset{k=1,\dots,K}{\operatorname{argmax}} \frac{\pi_k \cdot f_k(x)}{\sum_{j=1}^{k} \pi_j \cdot f_j(x)}}$$

papergrid

Date: / /

- **Classifiers and Discriminant Functions**

→ classifier $\delta = \delta(x) = \underset{k=1,\dots,K}{\operatorname{argmax}}\, g(x,k)$

$g = g(x,k)$ ↓ discriminant function

1. $g(x,k) = f(x) \cdot p(k|x) = f(x,k) = \pi_k \cdot f_k(x)$

2. $g(x,k) = \log(f(x,k)) = \log(\pi_k) + \log(f_k(x))$

3. $g(x,k) = \log(p(k|x)) = \log(\pi_k \cdot f_k(x)) - \log\left(\sum_{j=1}^{K} \pi_j \cdot f_j(x)\right)$

$p(k|x)$ – posterior probability

→ example

Date: / /

- Bayes Optimal Classifier for Identical Covariance matrices

→ In case of identical covariance matrices $\Sigma_K = \Sigma$, the discriminant function corresponding to the Bayes optimal classifier δ^* is given by

$$g(x,k) = \log(\pi_k) - \frac{1}{2}\log((2\pi)^D|\Sigma|) -$$

Since constant terms can be omitted, we get:

$$g(x,k) = \log(\pi_K) + x^T\Sigma^{-1}\mu_K - \frac{1}{2}\mu_K^T\Sigma^{-1}\mu_K$$

In special cases where $\Sigma_K = \Sigma$, $\underset{\text{disc. function}}{g(x,k)}$ is linear in x.

Q Exercise

$$f_K(x) = \frac{1}{(2\pi)^{3/2}|\Sigma_K|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma_K^{-1}(x-\mu_K)\right) ; k=1,2$$

with $\mu_1 = \begin{pmatrix}1\\2\\3\end{pmatrix}$ $\mu_2 = \begin{pmatrix}2\\2\\3\end{pmatrix}$ and $\Sigma_1 = \Sigma_2 = \Sigma = \begin{pmatrix}1&1&2\\1&9&4\\2&4&9\end{pmatrix}$

The observed feature vector should be $x = (1,2,2)^T$.
Determine the corresponding posterior probabilities $p(k|x)$ for the two cases:

i) $\pi_1 = \pi_2 = 0.5$ ii) $\pi_1 = 0.4$ $\pi_2 = 0.6$

$$p(k|x) = \frac{f(x,k)}{f(x)} = \frac{\pi_K \cdot f_K(x)}{\sum_{j=1}^{K}\pi_j \cdot f_j(x)}$$

$$\Sigma = \begin{pmatrix}1&1&2\\1&9&4\\2&4&9\end{pmatrix} \Rightarrow |\Sigma| = 36$$

$$\Sigma^{-1} = \frac{1}{36}\begin{pmatrix}65&-1&-14\\-1&5&-2\\-14&-2&8\end{pmatrix}$$

papergrid

Date: / /

Posterior Probabilities

$$P(k|x) = \frac{\pi_k \cdot f_k(\vec{x})}{f(x)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^{2} \pi_j f_j(x)}$$

$x = (1,2,2)^T$

$$= \frac{\pi_k \cdot \exp\left(-\frac{1}{2}(x-\mu_k)^T \cdot \Sigma^{-1}(x-\mu_k)\right)}{\pi_1 \cdot \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right) + \pi_2 \exp\left(-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)\right)}$$

Class 1 : $\exp\left(-\frac{1}{2}(0,2,-1) \cdot \frac{1}{36}\begin{pmatrix} 65 & -1 & -14 \\ -1 & 5 & -2 \\ -14 & -2 & 8 \end{pmatrix}\begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}\right) = \exp\left(-\frac{1}{2}\right)$

$= e^{-\frac{1}{2}}$

Class 2 : $\exp\left(-\frac{1}{2}(-1,0,-1) \cdot \frac{1}{36}\begin{pmatrix} 65 & -1 & -14 \\ -1 & 5 & -2 \\ -14 & -2 & 8 \end{pmatrix}\begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix}\right) = \exp\left(-\frac{5}{8}\right)$

$= e^{-5/8}$

i) $p(1|x) = \frac{0.5\, e^{-\frac{1}{2}}}{0.5 e^{-\frac{1}{2}} + 0.5 e^{-\frac{5}{8}}} \approx 0.531 \rightarrow \delta^*(x) = 1$

$x = (1,2,2)^T$

$p(2|x) = 1 - p(1|x) \approx 0.469 = \frac{0.5\, e^{-5/8}}{0.5 e^{-1/2} + 0.5 e^{-5/8}}$

ii) $p(1|x) = \frac{0.4\, e^{-\frac{1}{2}}}{0.4 e^{-\frac{1}{2}} + 0.6\, e^{-\frac{5}{8}}} \approx 0.430 \Rightarrow p(2|x) = ~~p~~\, 1 - p(1|x)$

≈ 0.570

$\hookrightarrow \delta^*(x) = 2$