# Bayes classifier (Bayes' decision rule)

Suppose there are M classes $C_1, C_2, \cdots, C_M$ ($M \geq 2$). Suppose the apriori prob. of the $i^{th}$ class is $P_i$, $i = 1, 2, \cdots, M$ and $\sum_{i=1}^{M} P_i = 1$.

[Apriori prob. of a class: If we pick any pattern from the whole set of patterns, the prob. of membership of that particular pattern (member) in that particular class]

Suppose the class cond$^{n}$al density fn for the $i^{th}$ class is denoted by $p_i$.

**Problem:** If $\underline{y}$ (observation vector / pattern) is the given obs$^{n}$ vector taken on a unit $v$, then how to put $v$ in one of the classes $C_1, C_2, \ldots, C_M$ based on $\underline{y}$.

**Example**

| | | | Apriori prob. | pdf |
|---|---|---|---|---|
| $M = 2$ | $C_1$ = Gujarati | 60% | 0.6 | $p_1$ |
| | $C_2$ = Punjabi | 40% | 0.4 | $p_2$ |

For each individual $v$ in a class we get a vector $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, say, and we get a big popl$^{n}$ of such vectors. Hence the prob. density fn. (pdf) is considered to be more or less alright.

For the present example let $M = 2$. Let $\Omega$ be the space of all possible obs$^{n}$s.

In the example above let $\underset{\sim}{y} = \begin{pmatrix} \text{ht.} \\ \text{wt.} \\ \vdots \end{pmatrix}$

Hence,

$$\Omega = (0, 200) \times (0, 150) \times (\cdots) \times (\cdots) \times \cdots$$

(in cm) (in kg)


Let $\Omega_1$ and $\Omega_2$ be s.t. $\Omega_1 \cup \Omega_2 = \Omega$

and $\Omega_1 \cap \Omega_2 = \phi$

Decision rule : If $\underset{\sim}{y} \in \Omega_1$ then put $v$ in $C_1$

If $\underset{\sim}{y} \in \Omega_2$ then put $v$ in $C_2$

**Question:** How to select $\Omega_1$ and $\Omega_2$ such that some optimality criterion is satisfied?

| Decision \ Reality | $v \in C_1$ | $v \in C_2$ |
|---|---|---|
| $\underline{y} \in \Omega_1$ ⇒ Put $v$ in $C_1$ | ✓ | * |
| $\underline{y} \in \Omega_2$ ⇒ Put $v$ in $C_2$ | * | ✓ |

"*" denotes misclassification

$$P(v \text{ is put in } C_2 \mid v \in C_1) = \int_{\Omega_2} p_1(\underline{x})\, d\underline{x}$$

$$P(v \text{ is put in } C_1 \mid v \in C_2) = \int_{\Omega_1} p_2(\underline{x})\, d\underline{x}$$

Hence the prob. of misclassification $= \epsilon(\Omega_1, \Omega_2)$

$= P(v \text{ is put in } C_2 \mid v \in C_1) \times P(v \in C_1)$

$+ P(v \text{ is put in } C_1 \mid v \in C_2) \times P(v \in C_2)$

$$= P_1 \int_{\Omega_2} p_1(\underline{x})\, d\underline{x} + P_2 \int_{\Omega_1} p_2(\underline{x})\, d\underline{x}$$

[Assumption: We know $p_1, p_2$ and $P_1, P_2$]

## Optimality Criterion

Choose $\Omega_1$ and $\Omega_2$ in such a way that minimizes $\epsilon(\Omega_1, \Omega_2)$.

$$\epsilon(\Omega_1, \Omega_2) = \int_{\Omega_2} P_1 p_1(\underline{x})\, d\underline{x} + \int_{\Omega_1} P_2 p_2(\underline{x})\, d\underline{x}.$$

$$\epsilon(\Omega_1, \Omega_2) = \int_{\Omega_2} P_1 p_1(\underline{x})\, d\underline{x} + \int_{\Omega_1} P_2 p_2(\underline{x})\, d\underline{x}$$

$$- \int_{\Omega_2} P_2 p_2(\underline{x})\, d\underline{x} + \int_{\Omega_2} P_2 p_2(\underline{x})\, d\underline{x}$$

$$= \int_{\Omega_2} \left( P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}) \right) d\underline{x} + \int_{\Omega_1 \cup \Omega_2} P_2 p_2(\underline{x})\, d\underline{x} .$$

$$= \int_{\Omega_2} \left( P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}) \right) d\underline{x} + P_2 \left[ \because \int_{\Omega_1 \cup \Omega_2} p_2(\underline{x})\, d\underline{x} = 1 \ (\text{pdf}) \right]$$

——— (1)

II$^{\text{nd}}$ly

$$\epsilon(\Omega_1, \Omega_2) = \int_{\Omega_1} \left( P_2 p_2(\underline{x}) - P_1 p_1(\underline{x}) \right) d\underline{x} + P_1$$

——— (2)

Adding ① and ② we get

$$2\epsilon(\Omega_1,\Omega_2) = P_1 + P_2 + \int_{\Omega_2} (P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}))\, d\underline{x}$$

$$+ \int_{\Omega_1} (P_2 p_2(\underline{x}) - P_1 p_1(\underline{x}))\, d\underline{x}$$

$$= 1 + \int_{\Omega_1} (P_2 p_2(\underline{x}) - P_1 p_1(\underline{x}))\, d\underline{x} + \int_{\Omega_2} (P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}))\, d\underline{x}$$

Minimization of $\epsilon(\Omega_1,\Omega_2)$ over $(\Omega_1,\Omega_2)$

$\Leftrightarrow$ Minimization of $2\epsilon(\Omega_1,\Omega_2)$ over $(\Omega_1,\Omega_2)$

$\Leftrightarrow$ Minimization of $\int_{\Omega_1} (P_2 p_2(\underline{x}) - P_1 p_1(\underline{x}))\, d\underline{x}$

$$+ \int_{\Omega_2} (P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}))\, d\underline{x}$$

over $(\Omega_1,\Omega_2)$

Let,

$$S_1 = \{ \underline{x} : P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}) > 0 \}$$

$$S_2 = \{ \underline{x} : P_1 p_1(\underline{x}) = P_2 p_2(\underline{x}) \}$$

$$S_3 = \{ \underline{x} : P_1 p_1(\underline{x}) - P_2 p_2(\underline{x}) < 0 \}$$

Note that $\Omega_1 \supseteq S_1$, $\Omega_2 \supseteq S_3$ and $S_2$ does not contribute in the minimization problem of misclassification. Thus, $S_2$ may become a subset of either $\Omega_1$ or $\Omega_2$. Thus without loss of generality, $\Omega_1 = S_1 \cup S_2$ and $\Omega_2 = S_3$ would actually minimize the prob. of misclassification.

## Remarks:

(1) Prob. of misclassification for $M = 2$ is $\leq \frac{1}{2}$

(2) For any $M$, the prob. of misclassification for Bayes' decision rule is $\leq \frac{M-1}{M}$.

(3) For an $M$ class classification problem, Bayes' decision rule provides $\Omega_i$ where

$$\Omega_i = \left\{ \underline{x} : P_i p_i(\underline{x}) \geq P_j p_j(\underline{x}), \forall i \neq j \right\}$$

## ASSN #1

Derive the Bayes' decision rule for $M = 3$

**Set up**: Let there be M classes $C_1, C_2, \ldots, C_M$ with apriori prob.s $p_1, p_2, \ldots, p_M$ and class cond$^{n}_{}$al prob.s (density fn$^{n}_{}$) $P_1, P_2, \ldots, P_M$. Let $\Omega$ be the space of all obs$^{n}_{}$s.

Let $S_1, S_2, \ldots, S_M$ be M sets such that

$\bigcup_{i=1}^{M} S_i = \Omega$ and $S_i \cap S_j = \phi, \forall i \neq j$.

Let a decision rule be derived as

$\underset{\sim}{y} \in S_i \Rightarrow$ the corresponding unit is placed in $C_i$.

With respect to this decision rule, $S_i$ is said to be the acceptance region for $C_i$.

The error prob for this decision rule is

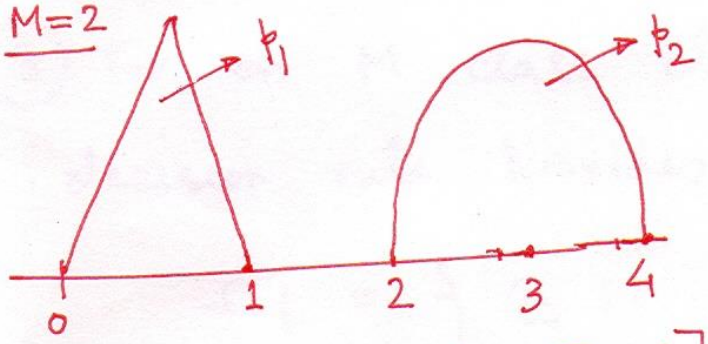
$$\epsilon(S_1, S_2, \ldots, S_M) = \sum_{i=1}^{M} p_i \int_{S_i^c} P_i(\underset{\sim}{x})\, d\underset{\sim}{x} \quad [S_i^c = \text{compliment of } S_i]$$

Let $\Omega_1, \Omega_2, \cdots, \Omega_M$ be the acceptance regions corresponding to Bayes rule, then

$$\epsilon(\Omega_1, \Omega_2, \cdots, \Omega_M) \leq \epsilon(S_1, S_2, \cdots, S_M), \quad \forall (S_1, S_2, \cdots, S_M)$$

## Example

$M=2$




$\Omega = [0,1] \cup [2,4]$

$P_1 = 0.3$
$P_2 = 0.7$

Let us take

$S_1 = [0, 1/2]$

$S_2 = (1/2, 1] \cup [2,4]$

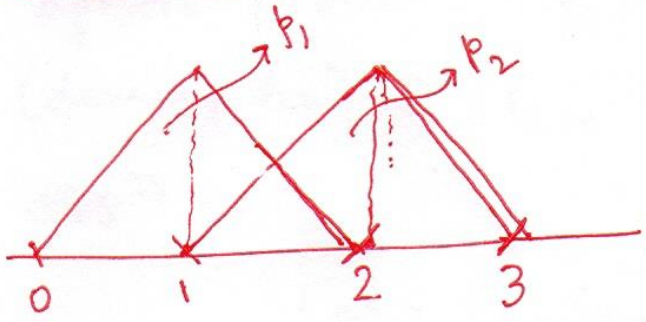
[We could have taken

$S_1 = [2,4]$ and $S_2 = [0,1]$

— It is another such pair.

We can construct many more such pairs to divide $\Omega$. Actually there are uncountably many such partition of $\Omega$.

## Example




$\Omega = [0, 3]$

Let $p_1(x) = \begin{cases} x, & 0 \le x \le 1 \\ 2-x, & 1 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$

$p_2(x) = \begin{cases} x-1, & 1 \le x \le 2 \\ 3-x, & 2 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$

Here, $p_1(x)$ and $p_2(x)$ are two density fns.

Let, $P_1 = 0.3$ and $P_2 = 0.7$.

(i) Let $S_1 = [0, 1]$ and $S_2 = (1, 3]$, then error prob. is

$$\epsilon(S_1, S_2) = \sum_{i=1}^{2} P_i \int_{S_i^c} p_i(x)\, dx$$

$$= P_1 \int_{(1,3]} p_1(x)\,dx + P_2 \int_{[0,1]} p_2(x)\,dx$$

$$= 0.3 \int_1^2 (2-x)\,dx + 0.7 \int_0^1 0 \cdot dx = 0.3 \left[2x - \frac{x^2}{2}\right]_1^2$$

$$= 0.3\left(2 - \frac{3}{2}\right) = \frac{0.3}{2} = 0.15.$$

(ii) Bayes decision rule :

$$\Omega_1 = \{ x : P_1 p_1(x) \geq P_2 p_2(x) \} \supseteq [0,1]$$

$$\Omega_2 = \{ x : P_2 p_2(x) \geq P_1 p_1(x) \} \supseteq [2,3]$$

Let $x \in [1,2]$; $P_1 p_1(x) \geq P_2 p_2(x) \Leftrightarrow 0.3(2-x) \geq 0.7(x-1)$

$\Leftrightarrow 0.6 - 0.3x \geq 0.7x - 0.7 \Leftrightarrow 1.3 \geq x$

Thus, $\Omega_1 = [0, 1] \cup [1, 1.3]$

and $\Omega_2 = (1.3, 2] \cup [2, 3]$

Error prob. for Bayes's decision rule is

$$= \epsilon(\Omega_1, \Omega_2) = \sum_{i=1}^{2} P_i \int_{\Omega_i^c} p_i(x)\, dx$$

$$= P_1 \int_{\Omega_2} p_1(x)\, dx + P_2 \int_{\Omega_1} p_2(x)\, dx$$

$$= 0.3 \int_{1.3}^{2} (2-x)\, dx + 0.7 \int_{1}^{1.3} (x-1)\, dx$$

$$= 0.3 \left[2x - \frac{x^2}{2}\right]_{1.3}^{2} + 0.7 \left[\frac{x^2}{2} - x\right]_{1}^{1.3}$$

$$= 0.042$$

The error prob. for Bayes decision rule is $\leq \frac{M-1}{M}$, where M is the number of classes.

**Proof**: Let the decision rule i be given as:

Put every unit in class i

$S_i = \Omega$ and $S_1 = S_2 = \cdots = S_{i-1} = S_{i+1} = \cdots = S_M = \phi$

The error prob. for the above rule

is $= \sum_{j=1}^{M} P_j \int_{S_j^c} p_j(x)\,dx$

$= \sum_{\substack{j=1 \\ (\neq i)}}^{M} P_j \int_{\Omega} p_j(x)\,dx + P_i \int_{\phi} p_i(x)\,dx$

$= \sum_{\substack{j=1 \\ (\neq i)}}^{M} P_j = 1 - P_i$

Let the error prob. for Bayes' decision rule be $\epsilon$

Then $\epsilon \leq 1 - P_i$, $\forall i = 1, 2, \cdots, M$

$$\text{or, } \epsilon \leq \min_{i=1,2,\cdots M} (1 - P_i) = 1 - \left( \max_{i=1,2,\cdots M} P_i \right)$$

$$\leq 1 - \frac{1}{M} = \frac{M-1}{M}$$

[ $\because \max_{i=1,2,\cdots M} P_i \geq \frac{1}{M}$

otherwise sum of M such quantities can not be 1

$\therefore - \max_{i=1,2,\cdots M} P_i \leq - \frac{1}{M}$ ]

Hence, the result.

# Bayes's Decision Rule under Normality Assumption

Let $p_i(\underline{x}) = \frac{1}{(\sqrt{2}\pi)^n |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)' \Sigma_i^{-1} (\underline{x}-\underline{\mu}_i)\right\}$

$i = 1, 2, \ldots, M$

$\Sigma_i$ is a $n \times n$ +ve definite matrix

[If A is +ve definite matrix then $\underline{x}' A \underline{x} \geq 0 \; \forall \underline{x} \neq \underline{0}$]

$|\Sigma_i| = \det(\Sigma_i)$

$\underline{\mu}_i$ = the mean vector of $c_i$

$\Sigma_i$ = the variance-co-variance of members of $c_i$.

$\underline{\mu}_i$ and $\Sigma_i$ are known.

Now

$P_i p_i(\underline{x}) \geq P_j p_j(\underline{x}), \forall j \neq i$

is resulting in

$$P_i \frac{1}{(\sqrt{2}\pi)^n |\Sigma_i|^{1/2}} \exp\left\{-\tfrac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right\}$$

$$\geq P_j \frac{1}{(\sqrt{2}\pi)^n |\Sigma_j|^{1/2}} \exp\left\{-\tfrac{1}{2}(\underline{x}-\underline{\mu}_j)'\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j)\right\} \quad \forall j \neq i$$

$$\Leftrightarrow \log P_i - \tfrac{1}{2}\log|\Sigma_i| - \tfrac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)$$

$$\geq \log P_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\underline{x}-\underline{\mu}_j)\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j) \quad \forall j \neq i$$

——— (I)

Case-1 Let $P_1 = P_2 = \cdots = P_M = \frac{1}{M}$ and $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_M = \Sigma$

(I) becomes

$$(\underline{x}-\underline{\mu}_i)'\Sigma^{-1}(\underline{x}-\underline{\mu}_i) \leq (\underline{x}-\underline{\mu}_j)'\Sigma^{-1}(\underline{x}-\underline{\mu}_j) \quad \forall j \neq i$$

——— (II)

The classifier (II) is known as Mahalanobis distance classifier

**Def$^{n}$:** Let there be two popl$^{n}$s with mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$ respectively and common variance co-variance matrix $\Sigma$. Then the Mahalanobis distance bet$^{n}$ the two popl$^{n}$s is denoted by $\Delta^2$ and is defined as

$$\Delta^2 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2).$$

**Case-2** Let $P_1 = P_2 = \cdots\cdots = P_M = \frac{1}{M}$

and $\Sigma_1 = \Sigma_2 = \cdots\cdots = \Sigma_M = I$

Then (I) becomes

$$(\underline{x} - \underline{\mu}_i)' (\underline{x} - \underline{\mu}_i) \leq (\underline{x} - \underline{\mu}_j)' (\underline{x} - \underline{\mu}_j) \quad \forall j \neq i$$

——————— (III)

**Case-3** $M = 2$, $\Sigma_1 = \Sigma_2 = \Sigma$

Now $P_1 p_1(\underline{x}) \geq P_2 p_2(\underline{x})$

$$\Leftrightarrow \log P_1 - \log P_2 \geq \frac{1}{2}\{(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1) - (\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\}$$

$$\Leftrightarrow \frac{1}{2}\{\underline{x}'\Sigma^{-1}\underline{x} - \underline{x}'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_1'\Sigma^{-1}\underline{x} + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - (\underline{x}'\Sigma^{-1}\underline{x} - \underline{\mu}_2'\Sigma^{-1}\underline{x} - \underline{x}'\Sigma^{-1}\underline{\mu}_2 + \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2)\} \leq \log\frac{P_1}{P_2}.$$

$$\Leftrightarrow \frac{1}{2}\{-2\underline{x}'\Sigma^{-1}\underline{\mu}_1 + 2\underline{x}'\Sigma^{-1}\underline{\mu}_2 + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2\} \leq \log\frac{P_1}{P_2}$$

$$\Leftrightarrow \underline{x}'\Sigma^{-1}(\underline{\mu}_2-\underline{\mu}_1) + \frac{1}{2}(\underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2) \leq \log\frac{P_1}{P_2} \quad \text{(IV)}$$

Thus, by Baye's rule the decision boundary bet$^{n}$ two classes is given by (IV) which is linear in $\underline{x}$. This is called a linear discriminant fn$^{}$ separating two classes.

- It can be shown that (IV) can be expressed as a fn$^{}$ of $\Delta^2$.

**Case – 4** $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_M = \Sigma$

Here the discriminatory fns are piecewise linear.

**Remarks**: 1. We may not know that the class cond$^{n}$al density fns are Normal.

2. Even if we know, it is difficult to find the exact value of Baye's error probability for any arbitrary class cond$^{n}$al density fn. In these case some bounds such as Chernoff bound, Bhattacharya bound could be used.

3. Some times aprori probabilities are hard to observe.

**Example**: Satellite image analysis where each pixel has 4 bands viz. blue band, green band, red band and infrared.

So, each pixel is a 4×1 vector and we need to classify each pixel in any one of the classes viz. water, veg., openspace and concrete.

Here it is very hard to observe the class prior prob.s.

## Estimation of parameters :

The parameters involve in the above cases are $\underline{\mu}$ and $\Sigma$.

$\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ are the given obs$^{n}$.s where

$$\underline{x}_l = (x_{l1}\ x_{l2}\ \ldots\ x_{ln})'$$

$$\underline{\bar{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix}$$

$$\bar{x}_k = \frac{1}{m} \sum_{l=1}^{m} x_{lk}$$

$k = 1, 2, \ldots, n$

Some times we need $\underline{\bar{x}}$ to normalize the obs$^{n}$.s.

Now out of "m" obs$^{n}$s, some will have class i (i = 1(1)M). Let the no. of obs$^{n}$ for each class is known. Then for each class we can observe mean (sample) vector $\bar{x}_i$ (i = 1(1)M) for each class.

[Note: From this inf. Sometimes $p_i$'s are also estimated]

## Estimation of $\Sigma$

$$\hat{\Sigma} = ((s_{ij}))_{n \times n}$$

where, $$s_{ij} = \frac{1}{m} \sum_{k=1}^{m} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

## Estimation of Error prob.

Let according to a decision rule $D_1$, $S_i$ be the acceptance region for $C_i$. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ be a sequence of obs$^{n}$s which need to be classified and the classification is known to us.

Let $y_i = \begin{cases} 1, & \text{if } \underline{x}_i \text{ is misclassified} \\ 0, & \text{otherwise.} \end{cases}$

Let $\bar{y}_m = \frac{1}{m} \sum_{i=1}^{m} y_i$

It can be shown that $\bar{y}_m$ "goes to" the error prob. of the decision rule $D_1$ as $m \to \infty$. $\bar{y}_m$ is called misclassification rate.

Example $M=2$, $p_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$

$$p_2(x) = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

Let $\mu_1 < \mu_2$

$P_1 = P_2 = \frac{1}{2}$

and $\sigma_1 = \sigma_2 = \sigma$

Now

$$P_1 p_1(x) > P_2 p_2(x)$$

$$\Leftrightarrow P_1 \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} > P_2 \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

$$\Leftrightarrow -\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 < -\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \quad \left[\begin{matrix} P_1 = P_2 \\ \sigma_1 = \sigma_2 \end{matrix}\right]$$

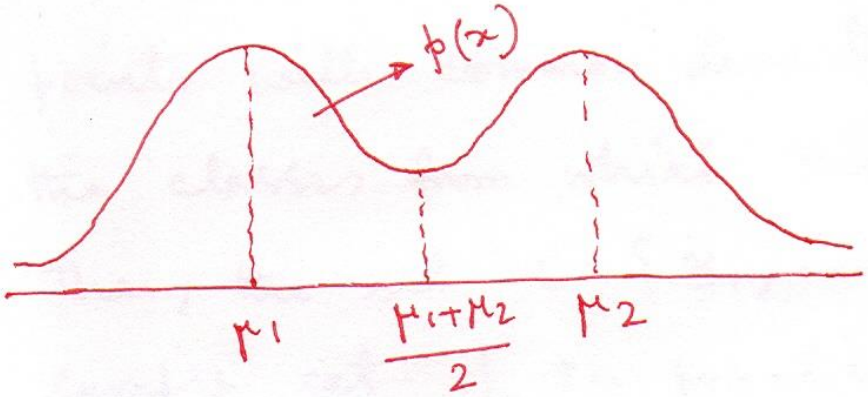
$$\Leftrightarrow (x-\mu_1)^2 < (x-\mu_2)^2$$

$$\Leftrightarrow x < (\mu_1 + \mu_2)/2.$$

- Mixture density fn., denoted by $p(x)$, is defined as

$$p(x) = \sum_{i=1}^{M} P_i\, p_i(x) \quad [\text{We have M classes}]$$

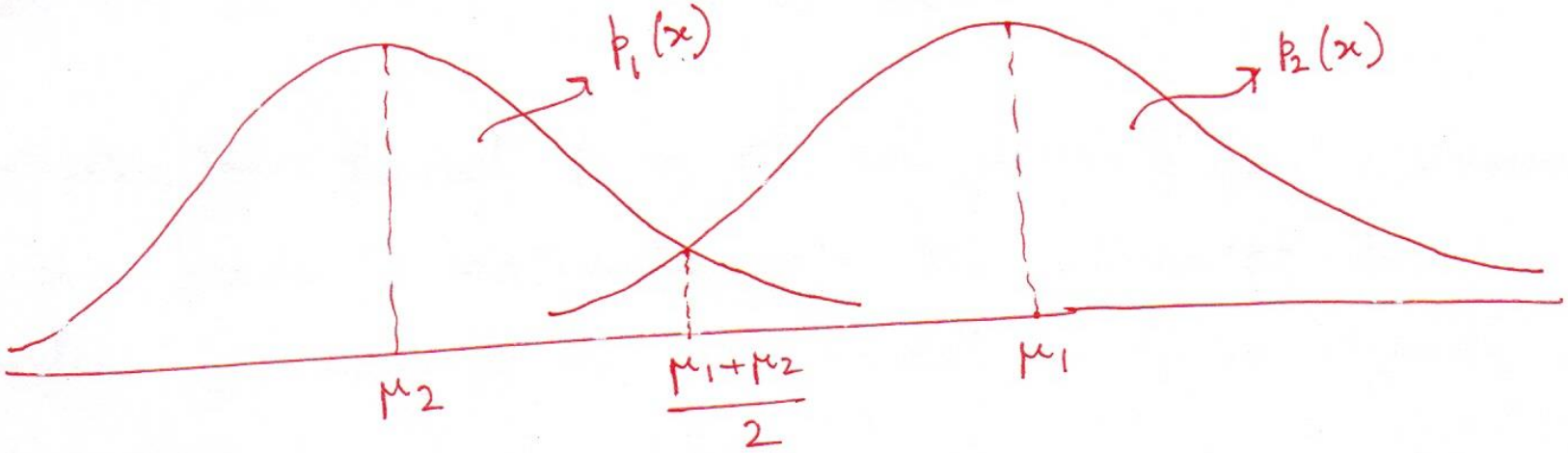
In the present problem $p(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\left\{e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} + e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2}\right\}$




The curve will look like, so, at $\frac{\mu_1+\mu_2}{2}$, the two curves $p_1(x)$ and $p_2(x)$ will meet

- Assumptions in the above example are: The apriori prob.s are same, the class density fn.s are Normal and the variance of these Normal dist.ns are equal.

Thus, $x < \frac{\mu_1+\mu_2}{2}$ which gives the Baye's decision boundary between two classes is represented as the valley of the histogram.




## Training Sample Set :

Training sample set is supposed to represent the properties of whole sample. In practise, we hope that training sample set actually represent the properties of popln.

Let there be M classes with apriori prob.s $P_1, \cdots, P_M$ and class density fn.s $p_1, \cdots, p_M$.

Let $p(x) = \sum_{i=1}^{M} P_i p_i(x)$.

Let $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_m$ be m randomly generated points with common density fn. $p(x)$. Suppose we know the classes from which these obs$^{n}$.s are obtained. Then, the set $S = \{\underline{x}_1, \cdots, \underline{x}_m\}$ is called training sample set. [In practice, it is very hard to obtain such set S.]

[We are to generate pt.s. from $p = 0.3\,N(\underline{\mu}_1, \Sigma_1) + 0.4\,N(\underline{\mu}_2, \Sigma_2) + 0.3\,N(\underline{\mu}_3, \Sigma_3)$.

Generate random NO.s [0,1]

If $r \in [0, 0.3]$ then draw the pt. from $N(\underline{\mu}_1, \Sigma_1)$
$r \in (0.3, 0.7]$ " " " " " $N(\underline{\mu}_2, \Sigma_2)$
$r \in (0.7, 1]$ " " " " " $N(\underline{\mu}_3, \Sigma_3)$ ]

- Let from a training sample set $S = \{\underline{x}_1, \cdots, \underline{x}_m\}$, $m_i$ be the number of obs$^{n}$/s from class $i$ i.e. $\sum_{i=1}^{M} m_i = m$.

Then we can show that $\frac{m_i}{m}$ "goes to" $P_i$ as $m \to \infty$.

In fact we can take $\frac{m_i}{m}$ as the estimate of $P_i$.

- Suppose the f$^{n}$.al form of the density f$^{n}$ is known, then one needs to estimate only the values of the parameters [Example of this with Normal density has already been discussed].

- If the form of the density f$^{n}$ is unknown, then we have to estimate it. There are ways of estimating this f$^{n}$. We will discuss one such method called <u>k-nearest neighbours method</u>.

## K-nearest neighbour density estimation procedure. (Loftsgaarden 1965)

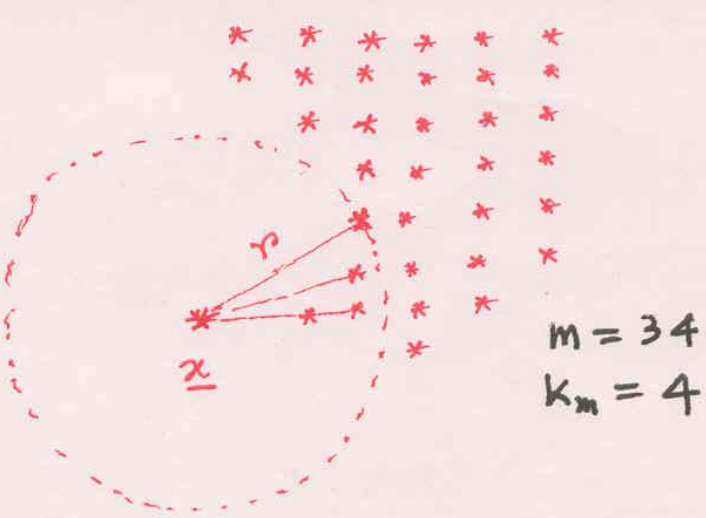
Let $f$ be a continous probability density fn. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m, \ldots$ be randomly generated pts. following the density fn. $f$ where $\underline{x}_i \in \mathbb{R}^n$, $i = 1(1)m \ldots$

Let $S_m = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m\}$. Let $\underline{x}$ be a point for which the density is to be estimated. Let $K_m$ be a positive number (integer) $< m$.

(i) Find $K_m$ nearest neighbours of $\underline{x}$ among the points in $S_m$. The $K_m^{\text{th}}$ nearest neighbour be at a distance $r$ from $\underline{x}$.

(ii) Let 'a' denotes the n-dimensional hyper volume of a sphere of radius 'r'.

Define $\hat{f}(x) = \frac{k_m}{m} \frac{1}{a}$.




It can be shown that,

$\hat{f}(x)$ "goes to" $f(x)$ as $m \to \infty$

If (i) $k_m \to \infty$ as $m \to \infty$ and

(ii) $\frac{k_m}{m} \to 0$ as $m \to \infty$

## K- nearest neighbour decision rule

Let there be M classes with class conditional densities $p_1, p_2, \ldots, p_M$ and apriori probabilities $P_1, P_2, \ldots, P_M$.

Let $p(\underline{x}) = \sum_{i=1}^{M} P_i p_i(\underline{x})$. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$, with $\underline{x}_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$ be $m$ randomly generated points with probability density $p(\underline{x})$.

Let $m_i$ of the $m$ points belong to class $i$ for all $i$

i.e. $\sum_{i=1}^{M} m_i = m$.

Let "K" be choosen properly. Let $\underline{x}$ be the point which needs to be classified using $\{\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_m\}$.

- Find K-nearest neighbours of $\underline{x}$ in the set $\{\underline{x}_1, \cdots, \underline{x}_m\}$.
- Let $K_i$ of these nearest neighbours belong to $i^{th}$ class,

  i.e. $\sum_{i=1}^{M} K_i = K$.
- Let the $K^{th}$ nearest neighbour lie at a distance $r$ from $\underline{x}$.
- Let "a" denotes the n-dimensional hyper volume of a sphere of radius $r$.

Now using nearest neighbour density estimation procedure, we have,

$$\left.\begin{aligned} \hat{p}(x) &= \frac{k}{m} \cdot \frac{1}{a} \\ \text{and } \hat{p}_i(x) &= \frac{k_i}{m_i} \cdot \frac{1}{a} \end{aligned}\right\}$$

Let $\hat{P}_i = \frac{m_i}{m}$.

Now use classifier rule with $\hat{P}_i$'s and $\hat{p}_i$'s.

So, classify the point $\underline{x}$ to the $i^{th}$ class

if $\hat{P}_i \hat{p}_i(\underline{x}) \geq \hat{P}_j \hat{p}_j(\underline{x}), \forall j \neq i$

$$\Leftrightarrow \frac{m_i}{m} \frac{k_i}{m_i} \frac{1}{a} \geq \frac{m_j}{m} \frac{k_j}{m_j} \frac{1}{a}, \forall j \neq i$$

$$\Leftrightarrow k_i \geq k_j, \forall j \neq i$$

[So, classify $\underline{x}$ in to $i^{th}$ class if $k_i = \max_{j=1(1)M} k_j$]

[We can remove the equality sign as we know from Bayes's decision rule that equality doesn't affect misclassification]

## Remarks :

1. K-NN decision rule is a non parametric decision rule in the sense that, the knowledge of apriori prob.s and class cond$^{n}$al densities are not necessary for this rule. We only assume that the given sample points are representatives of a mixture density fn.

2. It may so happen that one may be able to classify a point to more than one class.

If $K_{i_1} = K_{i_2} = \max_{i=1,\cdots M} K_i$, then the pattern may be classified to class $i_1$ or $i_2$.

In such a case one may increase the value of K by 1 and perform the classification task once again

3. When K=1, the rule is called "simply" nearest neighbour rule".

4. How to choose K ? — No satisfactory answer that can be applicable to all situations.

5. The K-nearest neighbour classificatein rule could be used to condens the training sample.

## Condensation of training sample set

**Algo for K=1**

Let $S = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m\}$ be the training data set (Note that class information for each data point is known).

(i) Keep one data point (randomly selected) in STORE

WLG Let STORE $= \{\underline{x}_1\}$

(ii) GRABBAG(0) = $S - \{\underline{x}_1\}$, GRABBAG(1) = $\phi$

(iii) Classify each $\underline{x}$ in GRABBAG(0) using K(=1 here) nearest neighbour rule with the help of points in STORE.

If $\underline{x}$ is misclassified, put $\underline{x}$ in STORE

otherwise put $\underline{x}$ in GRABBAG(1).

(iv) Write GRABBAG(1) as GRABBAG(0) and equate GRABBAG(1) to $\phi$.

(v) Repeat steps (iii) and (iv) till there are no transfer from GRABBAG(0) to STORE.

Remark: STORE represent the condensed training sample set.

Homework: Generalize the above algorithm to K-NN rule.

## Bayes decision rule for minimum risk

Let $M = 2$. Let the classes be denoted by $C_1$ and $C_2$.
Let $\Omega$ denotes the space of all possible observations. Let
$P_1, P_2$ denote the apriori prob.s of the classes and $p_1, p_2$
denote the class cond$^{n}_{n}$al density f$^{n}_{n}$s.

Let $a_{ij}$ denotes the cost (or loss) associated with putting an obs$^{n}_{n}$ into class $j$ given that it came from class $i$.

Let $a_{12} > a_{11} \geq 0$ and $a_{21} > a_{22} \geq 0$

Let $S_1$ and $S_2$ be sets such that $S_1 \cup S_2 = \Omega$ and $S_1 \cap S_2 = \phi$.

Let the decision rule be

$\underline{y} \in S_i \Rightarrow$ the corresponding unit $v$ is placed in class $i$.

Define $r$ = Expected cost (or loss).

## cost matrix

| | $v \in C_1$ | $v \in C_2$ |
|---|---|---|
| $\underline{y} \in S_1$ | $a_{11}$ | $a_{21}$ |
| $\underline{y} \in S_2$ | $a_{12}$ | $a_{22}$ |

## Expected cost

| | $v \in C_1$ | $v \in C_2$ |
|---|---|---|
| $\underline{y} \in S_1$ | $a_{11} P_1 \int_{S_1} p_1(x)\,dx$ | $a_{21} P_2 \int_{S_1} p_2(x)\,dx$ |
| $\underline{y} \in S_2$ | $a_{12} P_1 \int_{S_2} p_1(x)\,dx$ | $a_{22} P_2 \int_{S_2} p_2(x)\,dx$ |

$$P(y \in S_j \mid v \in C_i) = \int_{S_j} p_i(x)\,dx$$

$$\gamma = a_{11} P_1 P(\underline{y} \in S_1 \mid v \in C_1) + a_{12} P_1 P(y \in S_2 \mid v \in C_1)$$
$$+ a_{21} P_2 P(y \in S_1 \mid v \in C_2) + a_{22} P_2 P(y \in S_2 \mid v \in C_2)$$

$$= P_1 a_{12} \int_{S_2} p_1(x)\,dx + P_1 a_{11} \int_{S_1} p_1(x)\,dx + P_2 a_{22} \int_{S_2} p_2(x)\,dx + P_2 a_{21} \int_{S_1} p_2(x)\,dx$$

or,

$$\nu = P_1 a_{12} \int_{S_2} p_1(x)\,dx + P_1 a_{11} \int_{S_1} p_1(x)\,dx + P_1 a_{11} \int_{S_2} p_1(x)\,dx - P_1 a_{11} \int_{S_2} p_1(x)\,dx$$

$$+ P_2 a_{22} \int_{S_2} p_2(x)\,dx + P_2 a_{22} \int_{S_1} p_2(x)\,dx - P_2 a_{22} \int_{S_1} p_2(x)\,dx$$

$$+ P_2 a_{21} \int_{S_1} p_2(x)\,dx$$

or,

$$\nu = P_1 (a_{12} - a_{11}) \int_{S_2} p_1(x)\,dx + P_1 a_{11} \left( \int_{S_1} p_1(x)\,dx + \int_{S_2} p_1(x)\,dx \right)$$

$$+ P_2 (a_{21} - a_{22}) \int_{S_1} p_2(x)\,dx + P_2 a_{22} \left( \int_{S_2} p_2(x)\,dx + \int_{S_1} p_2(x)\,dx \right)$$

$$= P_1 (a_{12} - a_{11}) \int_{S_2} p_1(x)\,dx + P_2 (a_{21} - a_{22}) \int_{S_1} p_2(x)\,dx + P_1 a_{11} + P_2 a_{22}$$

——— (I)

Similarly by adjusting the terms of $\lambda$, we have

$$\lambda = P_1 a_{12} + P_2 a_{21} - P_1 (a_{12} - a_{11}) \int_{S_1} p_1(x)\,dx - P_2 (a_{21} - a_{22}) \int_{S_2} p_2(x)\,dx$$

(II)

Adding (I) and (II)

$$2\lambda = P_1 (a_{11} + a_{12}) + P_2 (a_{21} + a_{22}) + \int_{S_2} \{(a_{12} - a_{11}) P_1 p_1(x) - (a_{21} - a_{22}) P_2 p_2(x)\}\,dx + \int_{S_1} \{(a_{21} - a_{22}) P_2 p_2(x) - (a_{12} - a_{11}) P_1 p_1(x)\}\,dx$$

We would like to get $S_1$ and $S_2$ such that $\lambda$ is minimized.

Minimization of $\lambda \Longrightarrow$ Minimization of $2\lambda \Rightarrow$ Minimization of

$$\int_{S_1} \{(a_{21} - a_{22}) P_2 p_2(x) - (a_{12} - a_{11}) P_1 p_1(x)\}\,dx + \int_{S_2} \{(a_{12} - a_{11}) P_1 p_1(x) - (a_{21} - a_{22}) P_2 p_2(x)\}\,dx.$$

Now by examining terms of $2\bar{r}$, we define

$$\Omega_1 = \{ \underline{x} \mid (a_{21} - a_{22}) P_2 p_2(\underline{x}) - (a_{12} - a_{11}) P_1 p_1(\underline{x}) \leq 0 \}$$

and $\Omega_2 = \{ \underline{x} \mid (a_{12} - a_{11}) P_1 p_1(\underline{x}) - (a_{21} - a_{22}) P_2 p_2(\underline{x}) < 0 \}$

Then if $S_1 = \Omega_1$ and $S_2 = \Omega_2$, then we can minimize $\bar{r}$.

Note: $\{ \underline{x} \mid (a_{21} - a_{22}) P_2 p_2(\underline{x}) = (a_{12} - a_{11}) P_1 p_1(\underline{x}) \}$ is included in $\Omega_1$.

Remark: If $a_{12} - a_{11} = a_{21} - a_{22}$ then we obtain the Bayes decision rule.

Rule: $\Omega_1 = \{ \underline{x} \mid (a_{12} - a_{11}) P_1 p_1(\underline{x}) \geq (a_{21} - a_{22}) P_2 p_2(\underline{x}) \}$

$\Omega_2 = \{ \underline{x} \mid (a_{12} - a_{11}) P_1 p_1(\underline{x}) < (a_{21} - a_{22}) P_2 p_2(\underline{x}) \}$

- What are the acceptance region of Bayes decision rule for minimum risk when $M > 2$? (Home task)

# Kernel Density Estimation

# Estimation of Unknown PDF

- In many problems, underlying **pdf** has to be estimated from the available data.
- Sometimes, the type of the pdf (e.g. Gaussian) is known, but certain parameters such as mean and variance are not known.
- In contrast, sometimes parameters may be known but there is no information about the type of **pdf**.

- Depending on the available information, different approaches can be adopted.

# Approaches of Parametric Estimation

## Unknown Parameters and Known PDF

- Maximum Likelihood Parametric Estimation
- Maximum a Posteriori Probability Estimation
- Maximum Entropy Estimation
- The Expectation Maximization Algorithm

# Nonparametric Estimation

## Parameters may be known but PDF is unknown

- A kind of *histogram* approximation.
- The *x*-axis is first divided into successive bins of length h.
- Probability of sample x being located in a bin is estimated for each of the bins.
- If total number of samples = N, samples located in a bin = $k_N$, the corresponding probability is approximated by the frequency ratio $P \approx k_N/N$
- This approximation converges to actual P as N➔∞.
- The corresponding **pdf** is assumed to be constant throughout the bin and is approximated by

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_N}{N}, \quad |x - \hat{x}| \leq \frac{b}{2}$$

# Kernel Density Estimation (KDE)

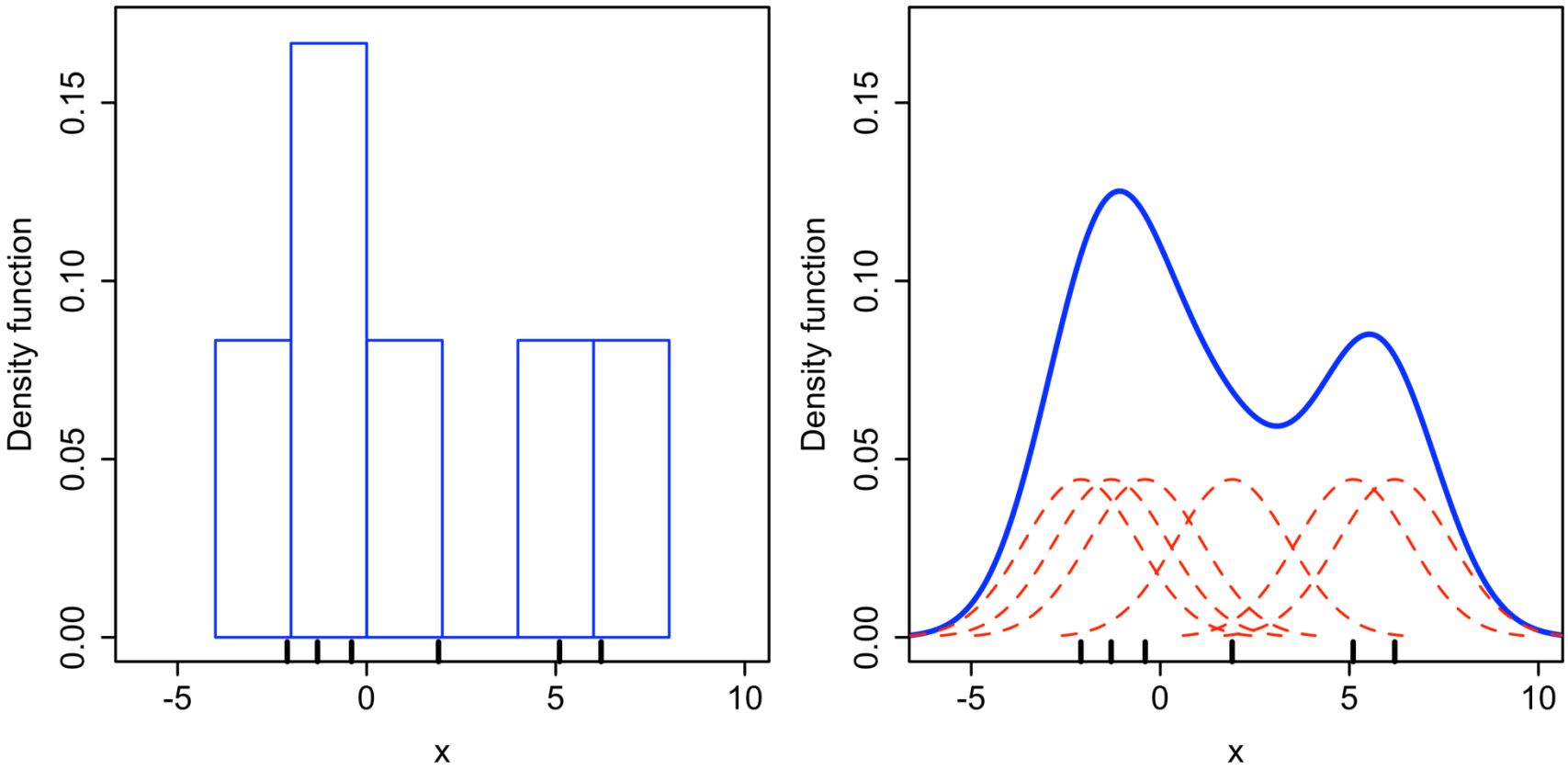
- KDE is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

- Method:
  - Let (x1, x2, …, xn) be an independently distributed samples drawn from some distribution with an unknown density function f.
  - We are interested in estimating the shape of this function f.
  - Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

where K() is the Kernel and h>0 is a smoothing parameter called the bandwidth.

Different Kernel functions used are: Uniform, Triangular, Tri-weight, Normal and others.




Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The 6 individual kernels are the red dashed curves, the kernel density estimate the blue curve. The data points are the rug plot on the horizontal axis.

- Algorithm

  - Let the given samples be x1, x2, …, xn.
  - Normalize the data, so that samples lie between 0 and 1.
  - Compute the range of this normalized data, i.e. [Min(xi), Max(xi)].
  - Take m points P = {p1, p2, …, pm} at equal interval within the computed range.
  - For every point pi, compute the probability density by summing up the densities calculated by the assumed (Gaussian) kernel with samples xj's as its mean. i.e.

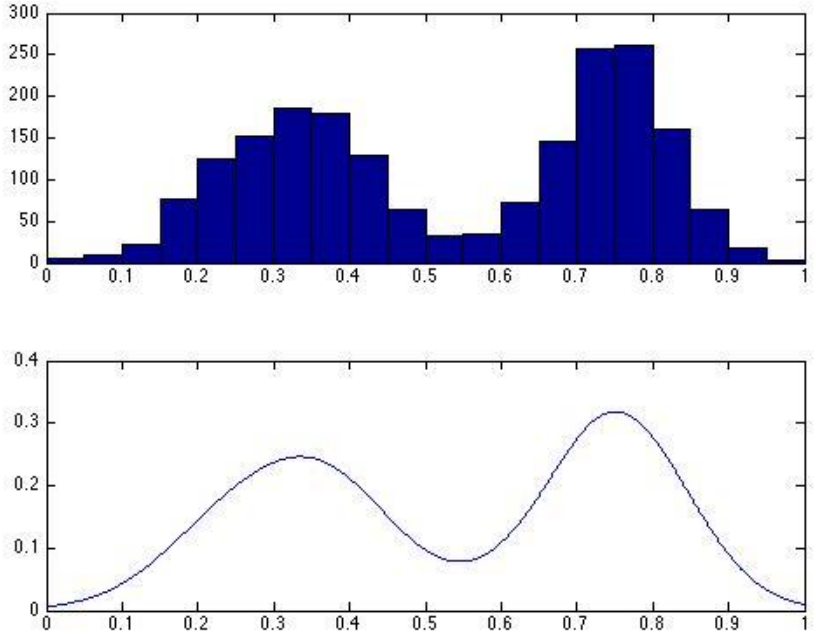
$$f(p_i) = \sum_{j=1}^{n} e^{-\frac{1}{2}\left(\frac{p_i - x_j}{\sigma}\right)^2}$$

- Plotting f against P, gives the required pdf.

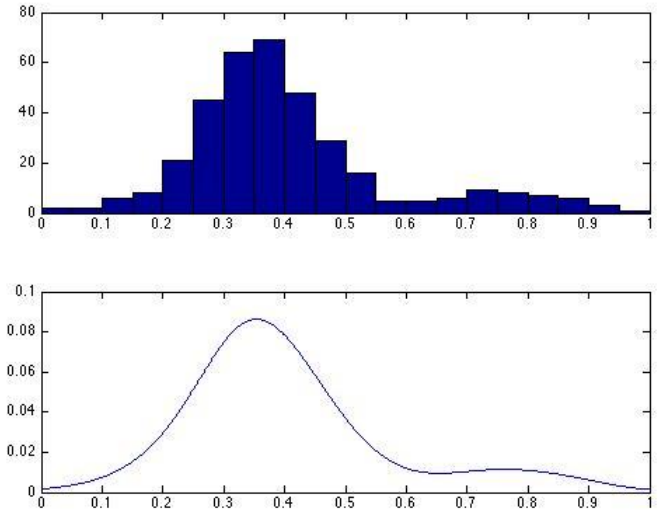
# Experiments

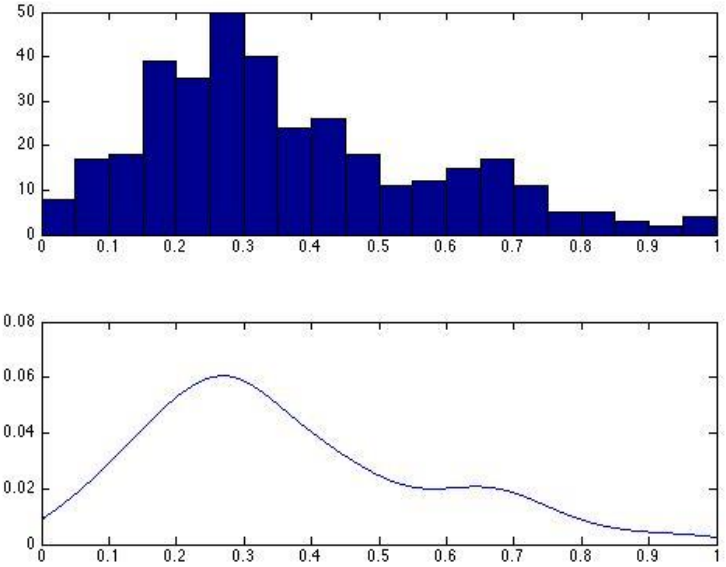
## Experiments 1

- Some experiments are performed taking h=1 and Gaussian Kernel function with standard deviation = 0.06
- First experiment draws some random samples from a Mixed Gaussian Distribution and computes the probability for 2000 points in the range of data.
- Second experiment takes the samples of a feature from LPP coefficients.



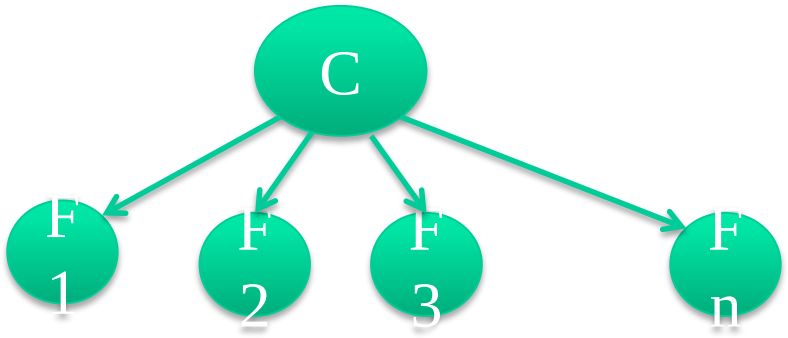
# Experiments 2



# Naïve Bayes Classifier

- NBC is a supervised Classifier.
- **Assumption:** Presence of a particular feature is independent to the presence of any other feature, given the class variable.
- **Structure**




- **Methodology:**
  - Since NBC assumes conditional independence of features, then joint model can be written as

$$p(C, F1, \ldots, Fn) = p(C)\prod_{i=1}^{n} p(Fi \mid C)$$

  - This means, that the conditional distribution over the class variable C can be expressed like this:

$$posterior = \frac{prior * likelihood}{evidence}$$

$$p(C \mid F1, \ldots, Fn) = \frac{p(C)\prod_{i=1}^{n} p(Fi \mid C)}{p(F1, \ldots, Fn)}$$

  - The class having highest probability calculated as above for a particular input pattern is assigned to it.