Pattern Classification

04. Bayes Classification Rule

AbdElMoniem Bayoumi, PhD

Fall 2021

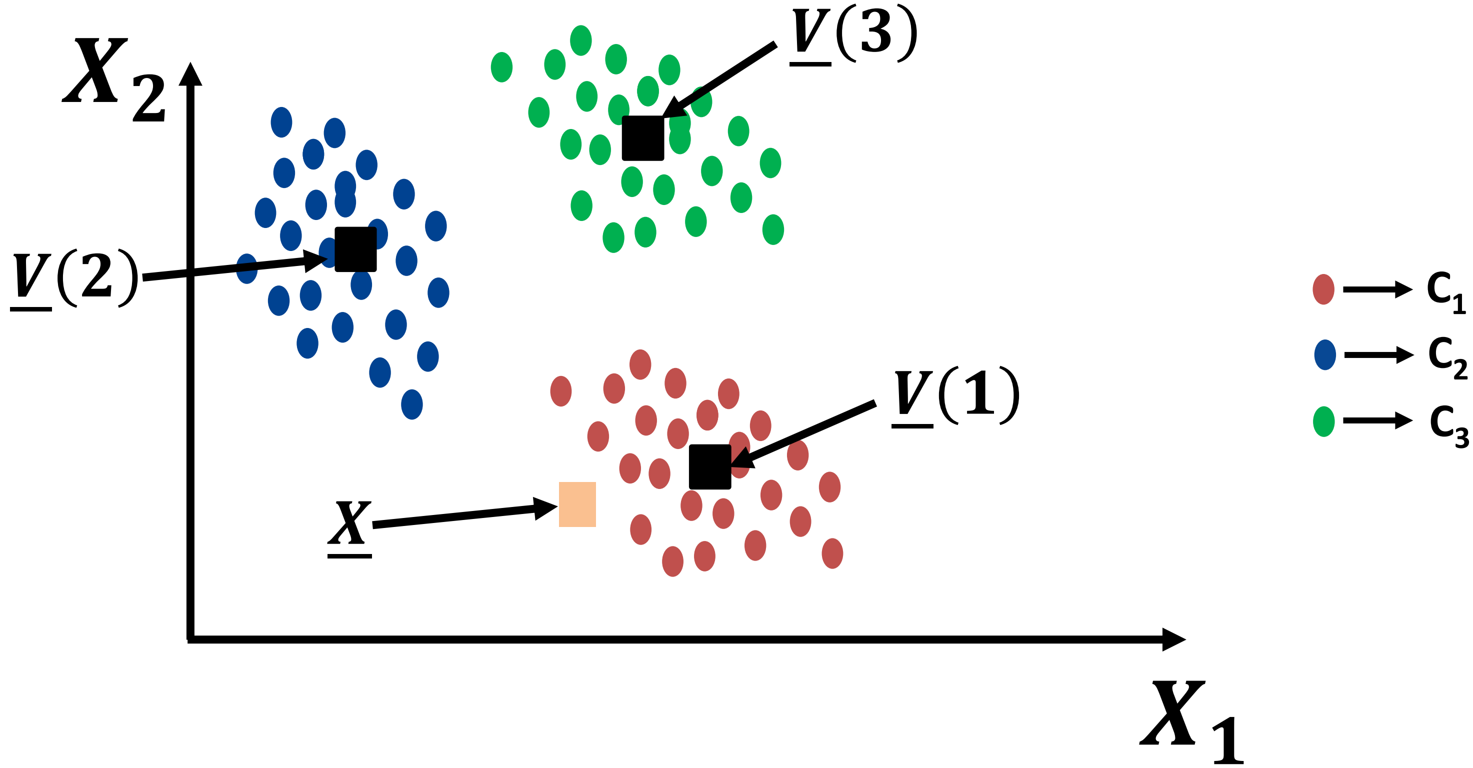
Acknowledgment

- These slides have been created relying on lecture notes of Prof. Dr. Amir Atiya

Recap: Minimum Distance Classifier

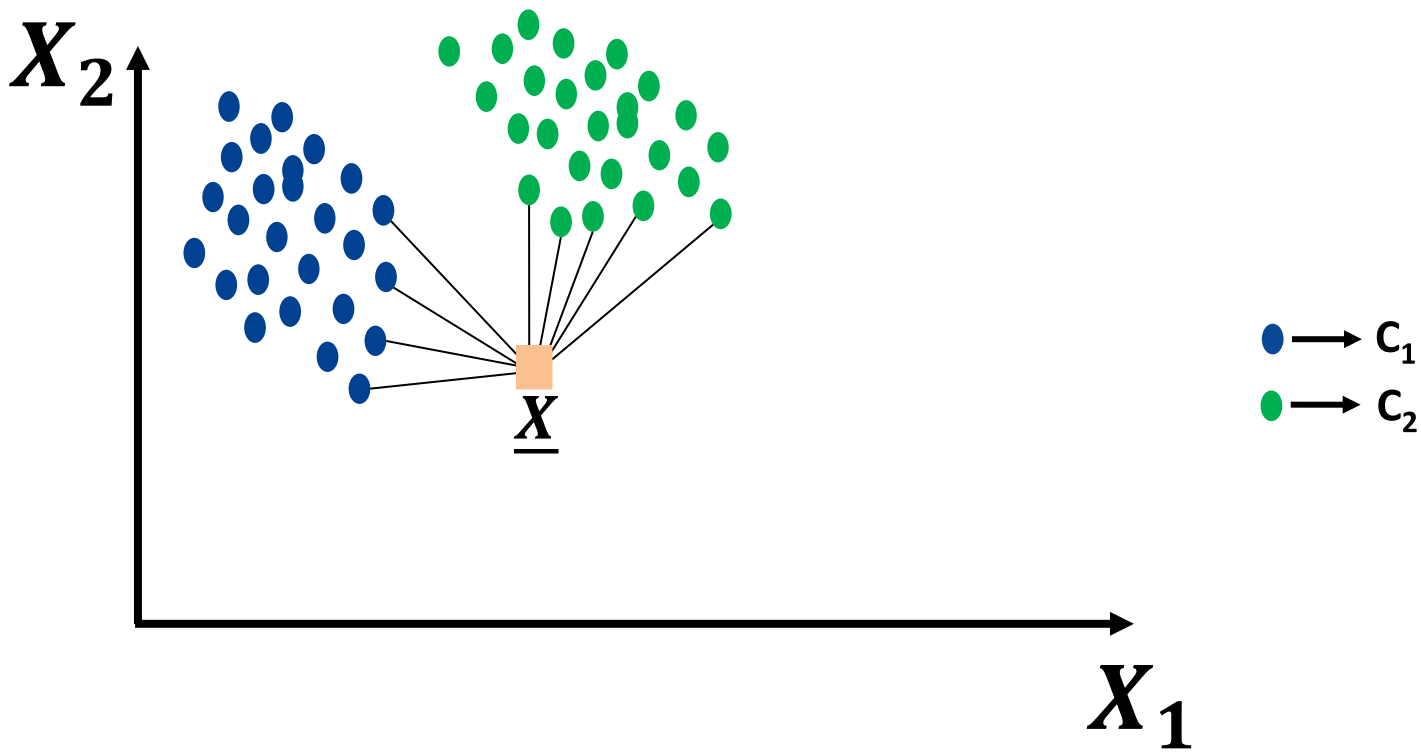
- Compute the distance from $\underline{X}$ to each center $\underline{V}(k)$:

$$\boldsymbol{d}(\boldsymbol{k}) = \sum_{i=1}^{N} [\boldsymbol{V}_i(\boldsymbol{k}) - \boldsymbol{X}_i]^2 \equiv \left\| \underline{\boldsymbol{V}}(\boldsymbol{k}) - \underline{\boldsymbol{X}} \right\|^2$$



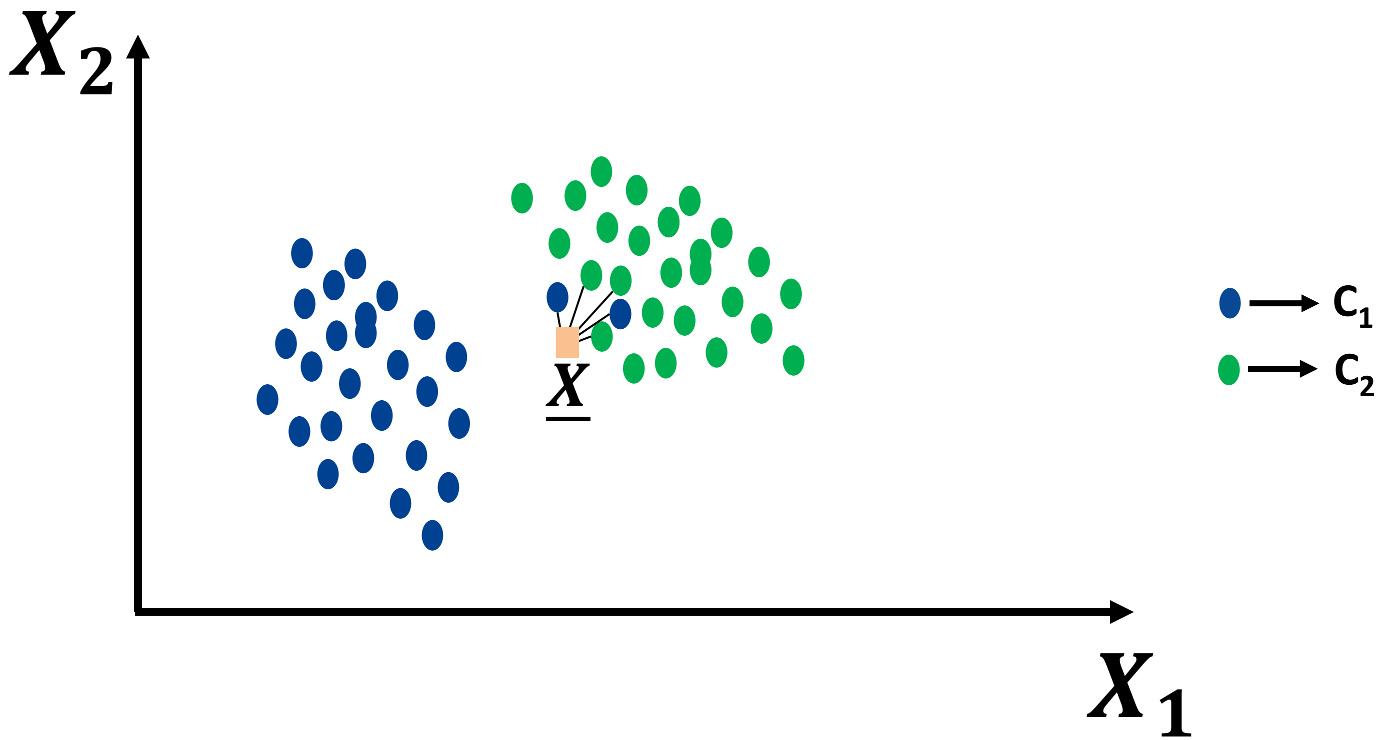
Recap: Nearest Neighbor Classifier

- The class of the nearest pattern to $\underline{X}$ determines its classification



Recap: K-Nearest Neighbor Classifier

- Take k = 5



- One can see that C_2 is the majority → classify $\underline{X}$ as C_2
- The KNN rule is less dependent on strange patterns compared to the nearest neighbor classification rule

Recap: Bayes Classification Rule

- To compute $P\left(C_i \middle| \underline{X}\right)$, we use Bayes rule:

$$P\left(C_i \middle| \underline{X}\right) = \frac{P(C_i\,, \underline{X})}{P(\underline{X})}$$

$$= \frac{P\left(\underline{X}|C_i\right) P(C_i)}{P(\underline{X})}$$

Bayes Rule:
P(A,B) = P(A|B)P(B) = P(B|A)P(A)

Recap: Bayes Classification Rule

- To compute $P(C_i|\underline{X})$, we use ***Bayes rule***:

$$P(C_i|\underline{X}) = \frac{P(\underline{X}|C_i)\,P(C_i)}{P(\underline{X})}$$

- $P(\underline{X}|C_i) \equiv$ Class-conditional density (***defined before***)

- $P(C_i) \equiv$ Probability of class C_i before or without observing the features $\underline{X}$
  $\equiv$ a priori probability of class C_i

Recap: Bayes Classification Rule

- The a priori probabilities represent the frequencies of the classes irrespective of the observed features

- For example in OCR, the a priori probabilities are taken as the frequency or fraction of occurrence of the different letters in a typical text

  – For the letters E & A → $P(C_i)$ will be higher

  – For letters Q & X → $P(C_i)$ will be low because they are infrequent

Bayes Classification Rule

- Find C_k giving max $P\left(C_k \middle| \underline{X}\right)$

$$P\left(C_k \middle| \underline{X}\right) = \frac{P\left(\underline{X} | C_k\right) P(C_k)}{P(\underline{X})}$$

  - $P\left(C_k \middle| \underline{X}\right) \equiv$ posterior prob.
  - $P(C_k) \equiv$ a priori prob.
  - $P\left(\underline{X} | C_k\right) \equiv$ class-conditional densities

- $P\left(\underline{X}\right) = \sum_{i=1}^{K} P(\underline{X}, C_i) = \sum_{i=1}^{K} P\left(\underline{X} \middle| C_i\right) P(C_i)$

Recap: Marginalization

- Discrete case:

$$P(A) = \sum_{i=1}^{N} P(A, B = B_i)$$

- Continuous case:

$$P(x) = \int_{-\infty}^{\infty} P(x, y)\, dy$$

- So:

$$P(\underline{X}) = \underbrace{\sum_{i=1}^{K} P(\underline{X}, C_i)}_{\text{Marginalization}} = \underbrace{\sum_{i=1}^{K} P(\underline{X} \mid C_i) P(C_i)}_{\text{Bayes rule}}$$

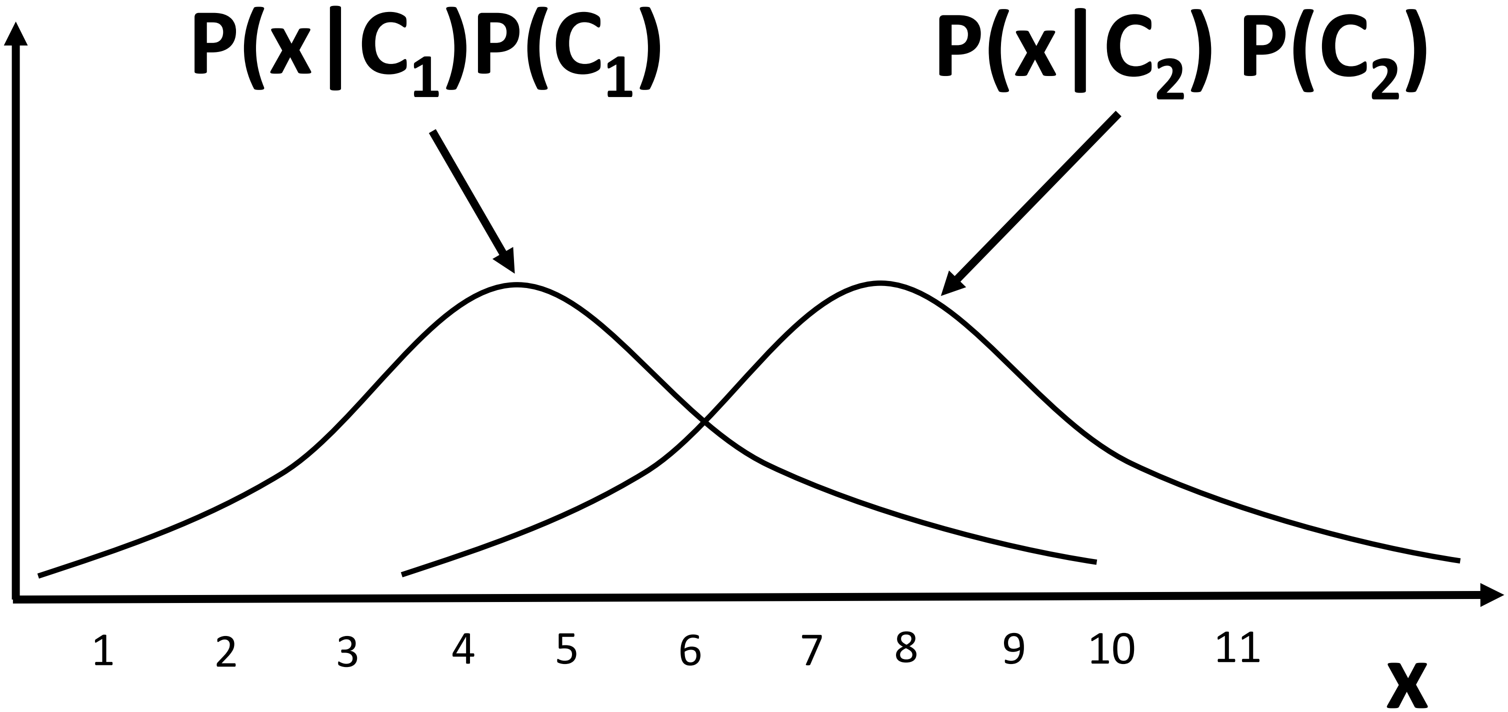
Bayes Classification Rule

$$P\left(C_k \middle| \underline{X}\right) = \frac{P\left(\underline{X}|C_k\right) P(C_k)}{\sum_{i=1}^{K} P\left(\underline{X} \middle| C_i\right) P(C_i)}$$

- In reality, we do not need to compute $P\left(\underline{X}\right)$ because it is a common factor for all the terms in the expression for $P\left(C_k \middle| \underline{X}\right)$
- Hence, it will not affect which terms will end up being maximum

Bayes Classification Rule

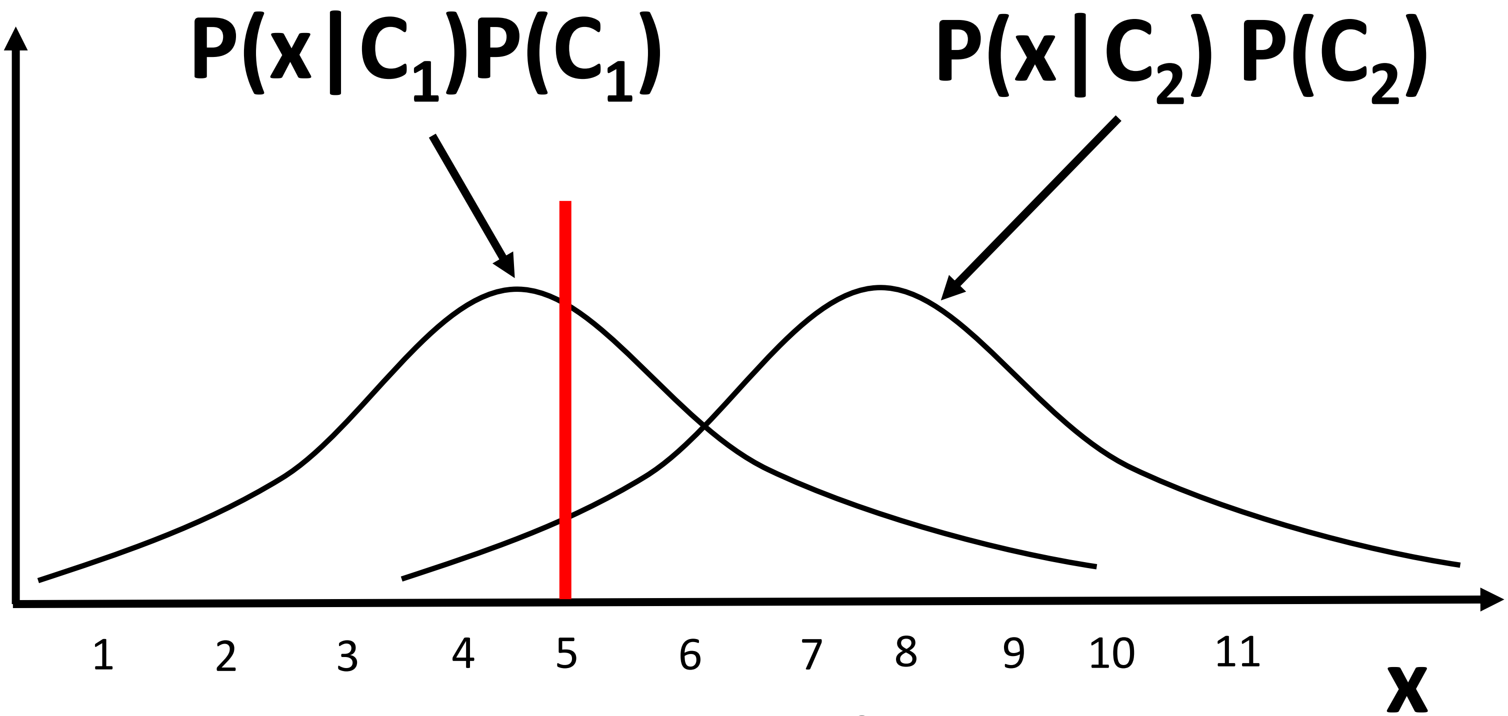
- Classify $\underline{X}$ to the class corresponding to max $P(\underline{X}|C_k)\,P(C_k)$



1-D example

Bayes Classification Rule

- Classify $\underline{X}$ to the class corresponding to max $P(\underline{X}|C_k)\,P(C_k)$



1-D example

- For x=5, $P(x|C_1)P(C_1)$ has a higher value compared to $P(x|C_2)P(C_2)$ → classify as C_1

Classification Accuracy

$$\boldsymbol{P\left(correct\ classification \middle| \underline{X}\right) = \max_{1 \le i \le K} P(C_i | \underline{X})}$$

- Example: 3-class case:
  - $P\left(C_1 \middle| \underline{X}\right) = 0.6$, $P\left(C_2 \middle| \underline{X}\right) = 0.3$, $P\left(C_3 \middle| \underline{X}\right) = 0.1$
  - You classified $\underline{X}$ as C_1 → it has highest $P(C_i | \underline{X})$
  - The probability that your classifier is correct equals to the probability that $\underline{X}$ belongs to the same class of the classification (which is 0.6)

Classification Accuracy

- Overall P(correct) is:

$$P(correct) = \int P(correct, \underline{X})\, d\underline{X}$$

Marginal prob.

$$= \int P(correct|\underline{X}) P(\underline{X})\, d\underline{X}$$

Bayes rule

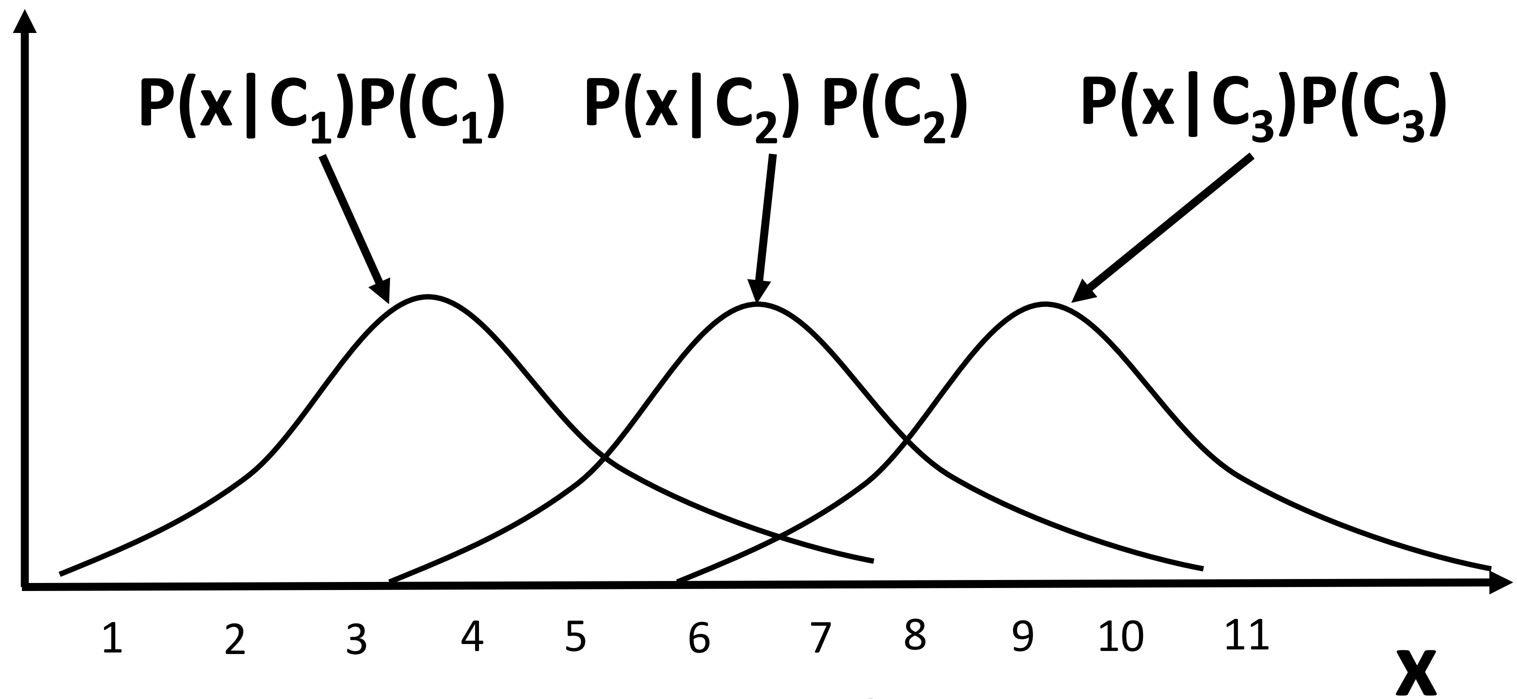
$$= \int \max_k \left[\frac{P(\underline{X}|C_k)\, P(C_k)}{\cancel{P(\underline{X})}} \right] \cancel{P(\underline{X})}\, d\underline{X}$$

$$= \int \max_k P(\underline{X}|C_k)\, P(C_k)\, d\underline{X}$$

Classification Accuracy

- Overall P(correct) is:

$$P(correct) = \int \max_k P(\underline{X}|C_k)\, P(C_k)\, d\underline{X}$$

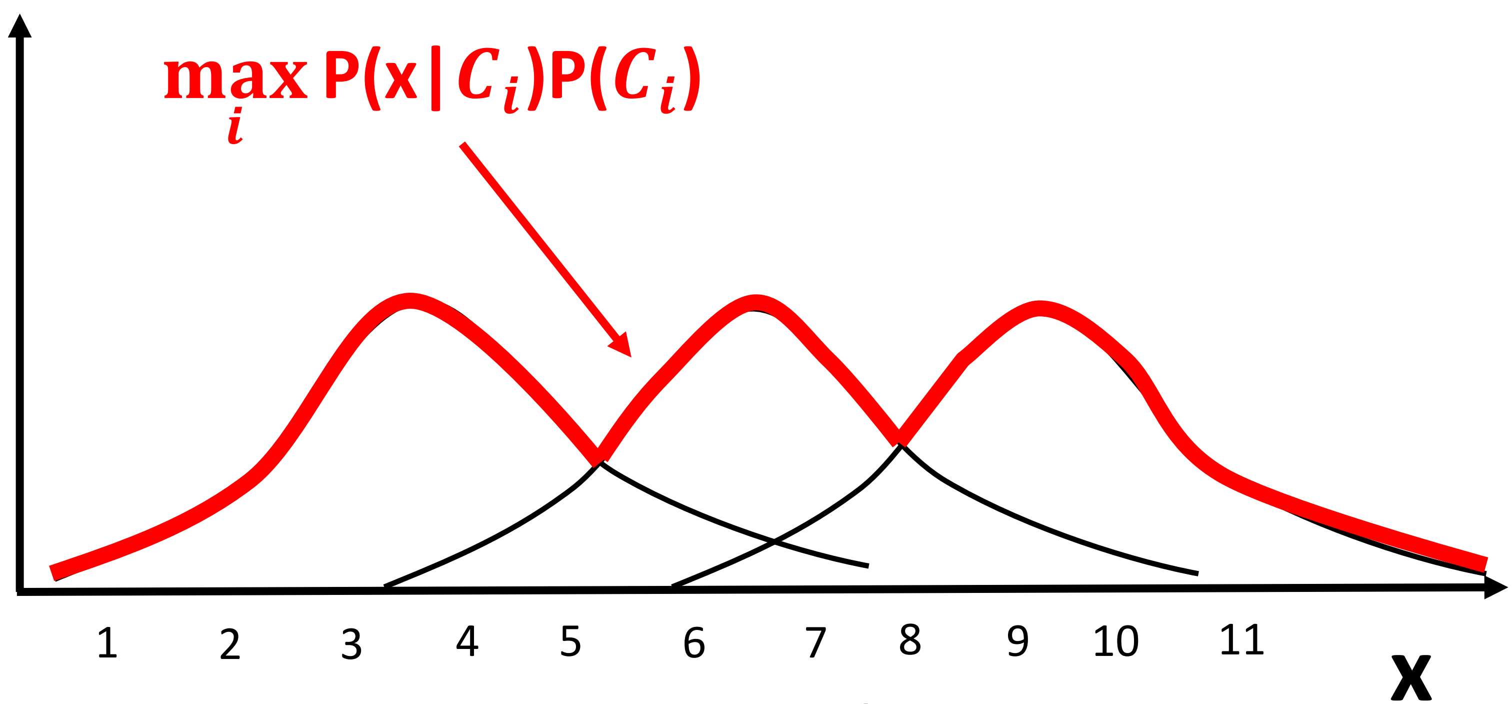


1-D example

Classification Accuracy

- Overall P(correct) is:

$$P(correct) = \int \max_k P(\underline{X}|C_k)\, P(C_k)\, d\underline{X}$$

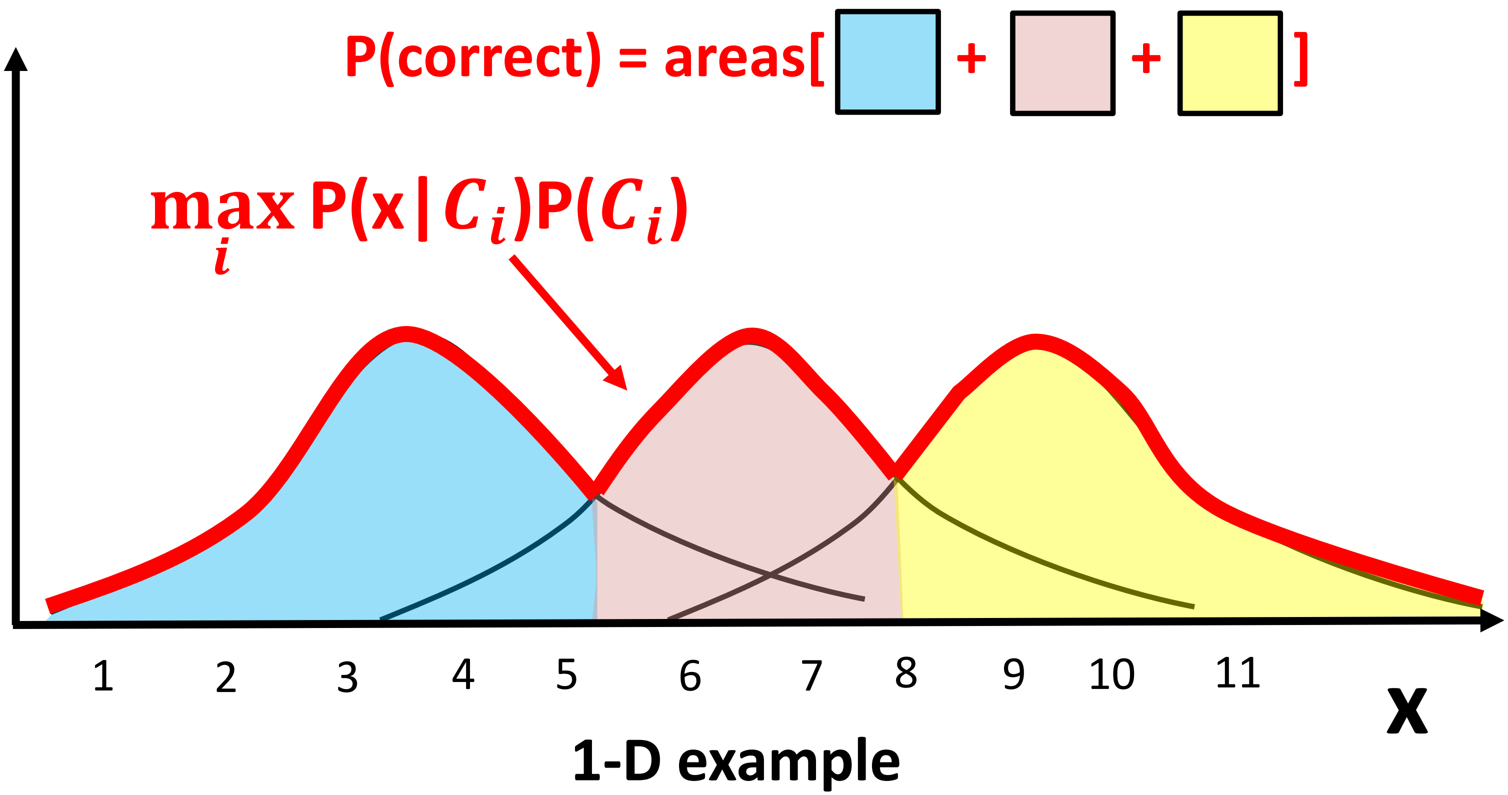


1-D example

Classification Accuracy

- Overall P(correct) is:

$$P(correct) = \int \max_k P(\underline{X}|C_k)\, P(C_k)\, d\underline{X}$$

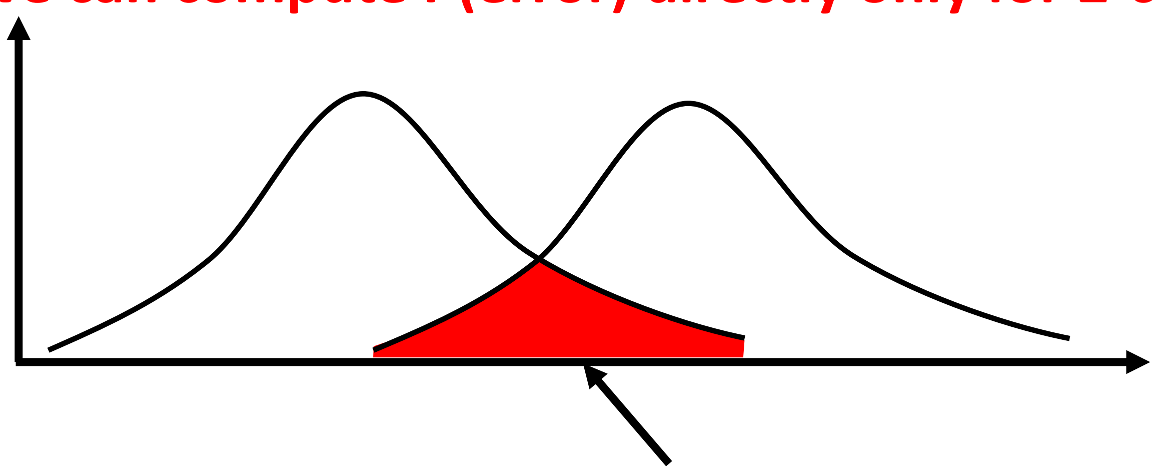


Classification Accuracy

$$P(correct) = \int \max_k P(\underline{X}|C_k)\, P(C_k)\, d\underline{X}$$

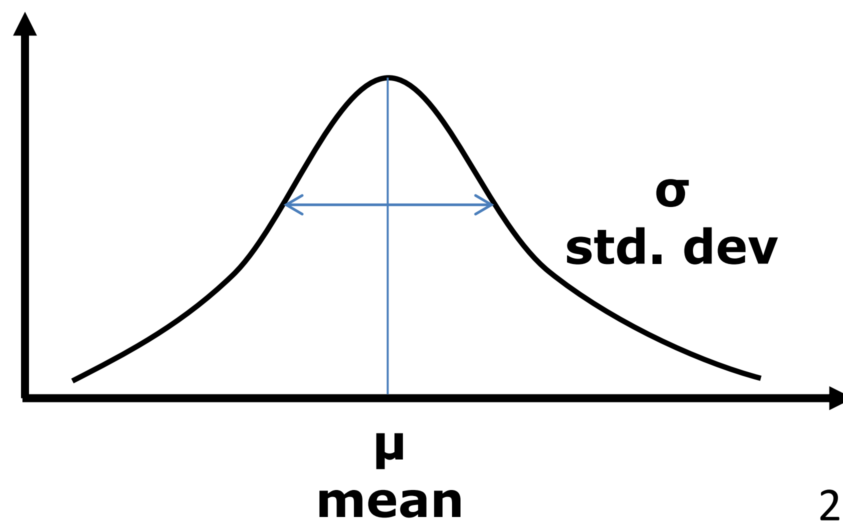
$$P(error) = 1 - P(correct)$$

We can compute P(error) directly only for 2-class case!



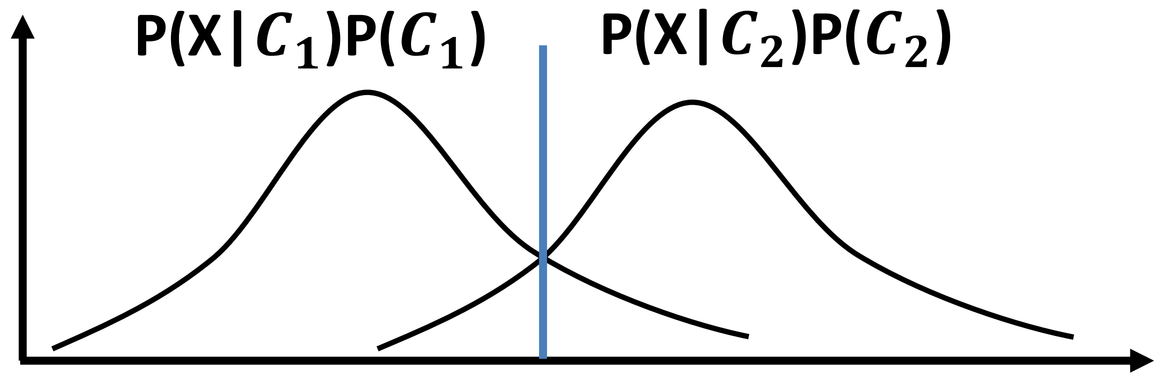
1-D Example

- Assume Gaussian class-conditional densities:
  - $P(X|C_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu_i)^2}{2\sigma^2}}$
  - $\mu_i = E(X) = \int \frac{X}{\sqrt{2\pi\sigma}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} dx$
  - $\text{Variance} = E[(X-\mu)^2] = \sigma^2$
  - $\sigma = \sqrt{var}$



1-D Example

- To get decision boundary:



$$P(X|C_1)P(C_1) = P(X|C_2)P(C_2)$$

1-D Example

- To get decision boundary:

$$P(X|C_1)P(C_1) = P(X|C_2)P(C_2)$$

$$P(C_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{\frac{-(X-\mu_1)^2}{2\sigma_1^2}} = P(C_2)\frac{1}{\sqrt{2\pi}\sigma_2}e^{\frac{-(X-\mu_2)^2}{2\sigma_2^2}}$$

- Take log of both sides ($log_e \equiv ln$), assuming $\sigma_1 = \sigma_2 = \sigma$:

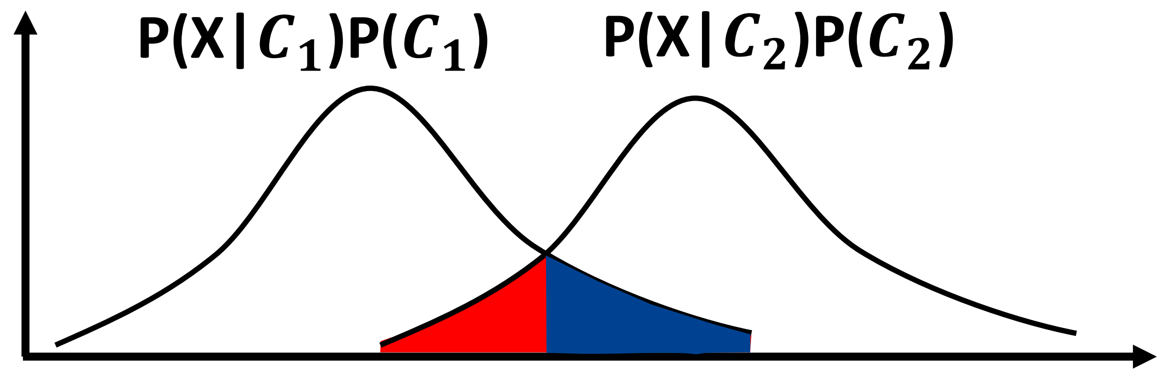
$$\log\left(P(C_1)\right) - \cancel{\log\left(\sqrt{2\pi}\sigma\right)} - \frac{(X-\mu_1)^2}{2\sigma^2} = log\left(P(C_2)\right) - \cancel{log\left(\sqrt{2\pi}\sigma\right)} - \frac{(X-\mu_2)^2}{2\sigma^2}$$

Can get $X_0 = fn(\mu_1, \mu_2, \sigma^2, P(C_1), P(C_2))$ → **quadratic equation**

Exercise!

1-D Example

- Compute P(error)



- $$P(error) = \underbrace{\int_{-\infty}^{X_0} \frac{P(C_2)e^{-\frac{(X-\mu_2)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX}_{I_1} + \underbrace{\int_{X_0}^{\infty} \frac{P(C_1)e^{-\frac{(X-\mu_1)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX}_{I_2}$$

1-D Example

- Compute P(error):

$$P(error) = \underbrace{\int_{-\infty}^{X_0} \frac{P(C_2)e^{-\frac{(X-\mu_2)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX}_{I_1} + \underbrace{\int_{X_0}^{\infty} \frac{P(C_1)e^{-\frac{(X-\mu_1)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX}_{I_2}$$

- Substitute in integral I_1:

$$\frac{X-\mu_2}{\sigma} = \acute{X}$$

$$\frac{dX}{\sigma} = d\acute{X}$$

Limits:

X = $-\infty \rightarrow \acute{X} = -\infty$

X = $X_0 \rightarrow \acute{X} = \frac{X_0-\mu_2}{\sigma}$

$$I_1 = P(C_2) \int_{-\infty}^{\frac{X_0-\mu_2}{\sigma}} \frac{e^{\frac{-\acute{X}^2}{2}}}{\sqrt{2\pi}} d\acute{X}$$

Normal integral CDF (numerically computed)

Do the same for I_2

Bayes Classification Rule

- It is an optimal classification rule
- This means no other classification rule is better (on average)
- The reason is that it chooses the most likely class (highest $P\left(C_i \middle| \underline{X}\right)$), so nothing could be better

this is a theoritical classification theorm, it may be applied if we have the value for the postirori probability, which is not always the case.

Bayes Classification Rule

- However, Bayes classifier assumes that the probability densities are known, which is not usually the case

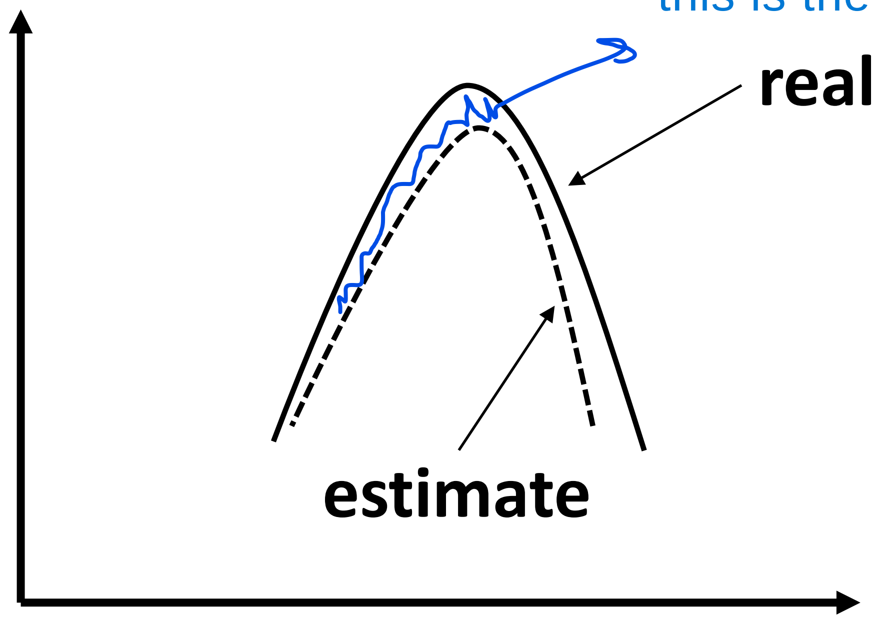
- Typically, only a training set is available and from the training set we can estimate the densities

- The density estimates have some "estimation error" that is higher if the training set size is smaller

l2n enta 34an t7sb el true value m7tag t3dy 3la el sha3b el masry kolo wahed wahed w t3ml 3leh el tagroba which is not logical y3ny, fa bn7sb el average.

Bayes Classification Rule

- Density estimation:



this is the error region

Gaussian Densities

- Assume a multi dimensional Gaussian density for each $P(\underline{X}|C_i)$
- Features may be independent (or conditionally independent), i.e., independent Gaussians
- Features may be dependent in other cases

the diffcult case is the dependent case, and this is the natural case, in real life a lot of things are dependent on each other.

Gaussian Densities (independent case)

- $P(\underline{X}|C_i) = P(X_1|C_i)P(X_2|C_i)\cdots P(X_N|C_i)$

  where $P(X_j|C_i) = \dfrac{e^{\frac{-(X_j-\mu_j)^2}{2\sigma_j^2}}}{\sqrt{2\pi}\sigma_j}$

- Get $P(\underline{X}|C_i) = \dfrac{e^{-\frac{1}{2}\sum_{j=1}^{N}\frac{(X_j-\mu_j)^2}{\sigma_j^2}}}{(2\pi)^{\frac{N}{2}}\sigma_1\sigma_2\cdots\sigma_N}$

Gaussian Densities (dependent case)

- $P\left(\underline{X} \middle| C_i\right) = \dfrac{e^{-\frac{1}{2}(\underline{X}-\underline{\mu})^T \Sigma^{-1} (\underline{X}-\underline{\mu})}}{(2\pi)^{\frac{N}{2}} det^{\frac{1}{2}}(\Sigma)}$

where:

Σ ≡ Covariance matrix (NxN)

det ≡ determinant

$\underline{\mu}$ ≡mean vector = $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$ $\qquad \mu_j = E(X_j)$

Gaussian Densities (dependent case)

- Let $Z = \underline{X} - \underline{\mu}$

- $Z^T A\, Z = Z_1^2 A_{11} + Z_1 Z_2 A_{12} + Z_1 Z_2 A_{21} + Z_2^2 A_{22} + Z_1 Z_3 A_{31} + \cdots$

Quadratic form

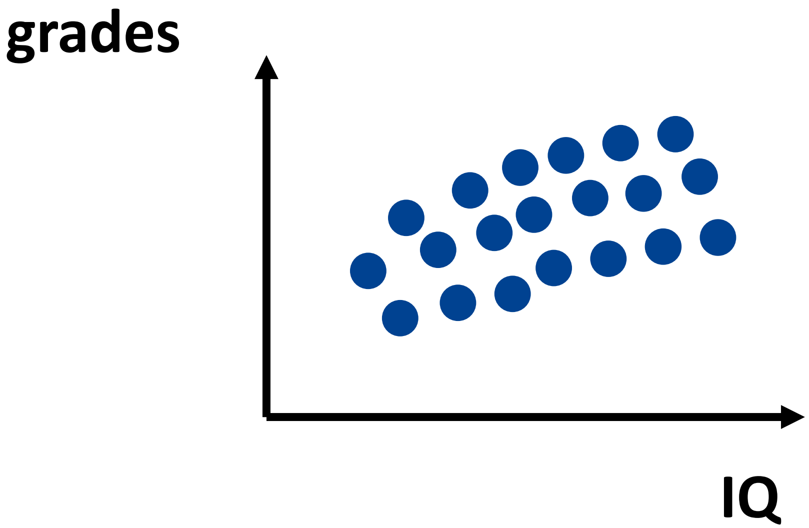
- A ≡ matrix

Exercise for yourself!

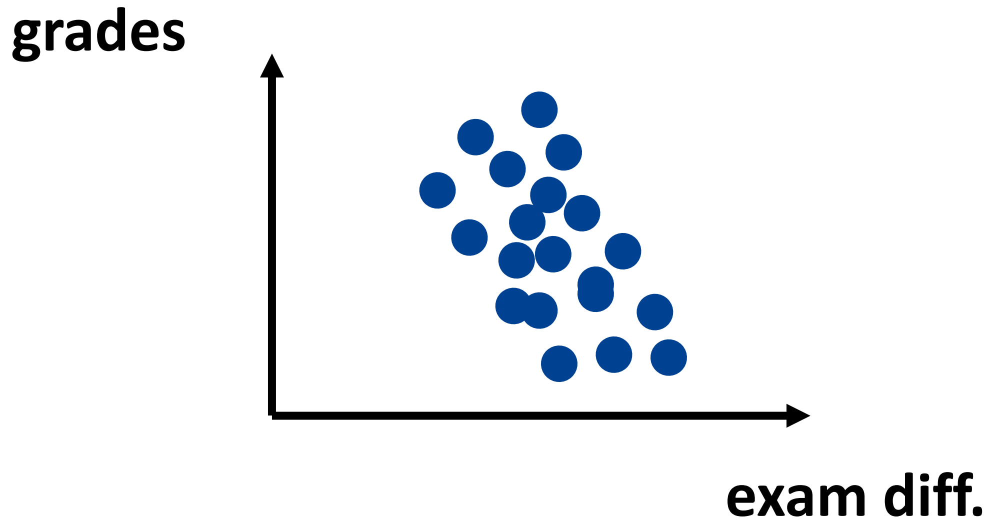
Covariance

- Covariance between two variables is a measure of how they change together
- Positive covariance if they tend to change in the same direction, e.g., weight & height
- Negative covariance if they tend to change in opposite directions, e.g., grade & exam difficulty
- Zero covariance if they are not related and do not influence each other's value

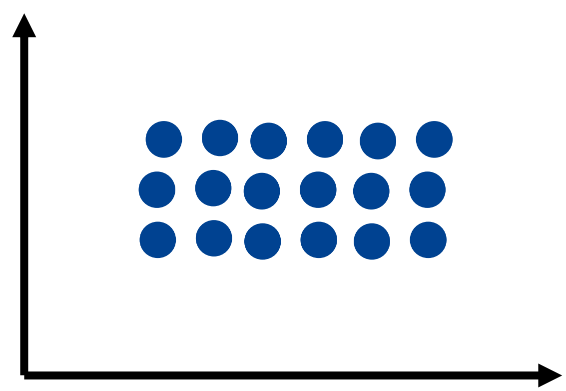
Covariance



positive covariance



negative covariance



zero covariance

Covariance Matrix

- Σ is a symmetric matrix, i.e., $\Sigma_{(j,k)} = \Sigma_{(k,j)}$
- $\Sigma_{(i,i)} = E[(X_i - \mu_i)^2] = \sigma_i^2$ → variance of ith item
- $\Sigma_{(j,k)} = E\left[\left(X_j - \mu_j\right)(X_k - \mu_k)\right] \equiv$ covariance(X_j, X_k)
- $\Sigma = \begin{bmatrix} E[(X_1-\mu_1)^2] & E[(X_1-\mu_1)(X_2-\mu_2)] & \cdots \\ E[(X_1-\mu_1)(X_2-\mu_2)] & E[(X_2-\mu_2)^2] & \vdots \\ \vdots & \vdots & \ddots \end{bmatrix}$

Some Properties

- If X is Gaussian then its linear transformation is also Gaussian

- For the independent case:

$$E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = 0 \qquad if\ i \neq j$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{bmatrix}$$ **diagonal matrix**

Some Properties

- For the independent case:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_N^2} \end{bmatrix}$$

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2 \cdots \sigma_N^2$$

$$\left(\underline{X} - \underline{\mu}\right)^T \Sigma^{-1} \left(\underline{X} - \underline{\mu}\right)$$

$$P\left(\underline{X} \middle| C_i\right) = \frac{e^{-\frac{1}{2} \sum_{j=1}^{N} \frac{(X_j - \mu_j)^2}{\sigma_j^2}}}{(2\pi)^{\frac{N}{2}} \sigma_1 \sigma_2 \cdots \sigma_N}$$

$$det^{\frac{1}{2}}(\Sigma)$$

$$\Sigma^{-1}\left(\underline{X} - \underline{\mu}\right) = \begin{bmatrix} \frac{X_1 - \mu_1}{\sigma_1^2} \\ \frac{X_2 - \mu_2}{\sigma_2^2} \\ \vdots \\ \frac{X_N - \mu_N}{\sigma_N^2} \end{bmatrix}$$

Dependent case:

$$P\left(\underline{X} \middle| C_i\right) = \frac{e^{-\frac{1}{2}(\underline{X}-\underline{\mu})^T \Sigma^{-1} (\underline{X}-\underline{\mu})}}{(2\pi)^{\frac{N}{2}} det^{\frac{1}{2}}(\Sigma)}$$

Decision boundaries

- Decision boundary in case of multi-dimensional Gaussian class-conditional densities:

$$P(C_i)P\left(\underline{X}\middle|C_i\right) = P\left(C_j\right)P\left(\underline{X}\middle|C_j\right)$$

$$P(C_i)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_i}\right)^T\Sigma_i^{-1}(\underline{X}-\underline{\mu_i})}}{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_i)} = P\left(C_j\right)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_j}\right)^T\Sigma_j^{-1}(\underline{X}-\underline{\mu_j})}}{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_j)}$$

Solve to get equation of boundary condition!

Decision boundaries

- For simplicity assume $\Sigma_i = \Sigma_j = \Sigma$

$$P(C_i)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu}_i\right)^T\Sigma_i^{-1}(\underline{X}-\underline{\mu}_i)}}{\cancel{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_i)}} = P(C_j)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu}_j\right)^T\Sigma_j^{-1}(\underline{X}-\underline{\mu}_j)}}{\cancel{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_j)}}$$

Decision boundaries

- For simplicity assume $\Sigma_i = \Sigma_j = \Sigma$

$$P(C_i)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_i}\right)^T \Sigma_i^{-1}(\underline{X}-\underline{\mu_i})}}{\cancel{(2\pi)^{\frac{N}{2}} det^{\frac{1}{2}}(\Sigma_i)}} = P(C_j)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_j}\right)^T \Sigma_j^{-1}(\underline{X}-\underline{\mu_j})}}{\cancel{(2\pi)^{\frac{N}{2}} det^{\frac{1}{2}}(\Sigma_j)}}$$

$$\log(P(C_i)) - \frac{1}{2}\left(\underline{X}-\underline{\mu_i}\right)^T \Sigma^{-1}\left(\underline{X}-\underline{\mu_i}\right) = \log\left(P(C_j)\right) - \frac{1}{2}\left(\underline{X}-\underline{\mu_j}\right)^T \Sigma^{-1}\left(\underline{X}-\underline{\mu_j}\right)$$

$$2\left[\log(P(C_i)) - \log\left(P(C_j)\right)\right]$$
$$= -\left[\cancel{\underline{X}^T\Sigma^{-1}\underline{X}} \underbrace{- \underline{X}^T\Sigma^{-1}\underline{\mu_j} - \underline{\mu_j}^T\Sigma^{-1}\underline{X}}_{-2\underline{\mu_j}^T\Sigma^{-1}\underline{X}} + \underline{\mu_j}^T\Sigma^{-1}\underline{\mu_j}\right] + \left[\cancel{\underline{X}^T\Sigma^{-1}\underline{X}} \underbrace{- \underline{X}^T\Sigma^{-1}\underline{\mu_i} - \underline{\mu_i}^T\Sigma^{-1}\underline{X}}_{-2\underline{\mu_i}^T\Sigma^{-1}\underline{X}} + \underline{\mu_i}^T\Sigma^{-1}\underline{\mu_i}\right]$$

$$\boldsymbol{\underline{X}^T\Sigma^{-1}\underline{\mu_j} = (\underline{X}^T\Sigma^{-1}\underline{\mu_j})^T = \underline{\mu_j}^T\Sigma^{-1}\underline{X}}$$

Scalar = Scalar ^{T}

$$\boldsymbol{(ABC)^T = C^T B^T A^T}$$

$$\boldsymbol{(\underline{X}^T\Sigma^{-1}\underline{\mu_j})^T = \underline{\mu_j}^T(\Sigma^{-1})^T\underline{X}}$$

$\boldsymbol{(\Sigma^{-1})^T = \Sigma^{-1}}$ **symmetric matrix**

Decision boundaries

- For simplicity assume $\Sigma_i = \Sigma_j = \Sigma$

$$P(C_i)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_i}\right)^T\Sigma_i^{-1}(\underline{X}-\underline{\mu_i})}}{\cancel{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_i)}} = P(C_j)\frac{e^{-\frac{1}{2}\left(\underline{X}-\underline{\mu_j}\right)^T\Sigma_j^{-1}(\underline{X}-\underline{\mu_j})}}{\cancel{(2\pi)^{\frac{N}{2}}det^{\frac{1}{2}}(\Sigma_j)}}$$

$$\log(P(C_i)) - \frac{1}{2}\left(\underline{X}-\underline{\mu_i}\right)^T\Sigma^{-1}\left(\underline{X}-\underline{\mu_i}\right) = \log\left(P(C_j)\right) - \frac{1}{2}\left(\underline{X}-\underline{\mu_j}\right)^T\Sigma^{-1}\left(\underline{X}-\underline{\mu_j}\right)$$

$$2\left[\log(P(C_i)) - \log\left(P(C_j)\right)\right]$$
$$= -\left[\cancel{\underline{X}^T\Sigma^{-1}\underline{X}} \underbrace{- \underline{X}^T\Sigma^{-1}\underline{\mu_j} - \underline{\mu_j}^T\Sigma^{-1}\underline{X}}_{-2\underline{\mu_j}^T\Sigma^{-1}\underline{X}} + \underline{\mu_j}^T\Sigma^{-1}\underline{\mu_j}\right] + \left[\cancel{\underline{X}^T\Sigma^{-1}\underline{X}} \underbrace{- \underline{X}^T\Sigma^{-1}\underline{\mu_i} - \underline{\mu_i}^T\Sigma^{-1}\underline{X}}_{-2\underline{\mu_i}^T\Sigma^{-1}\underline{X}} + \underline{\mu_i}^T\Sigma^{-1}\underline{\mu_i}\right]$$

$$2log\left[\frac{P(C_i)}{P(C_j)}\right] = 2\underline{\mu_j}^T\Sigma^{-1}\underline{X} - \underline{\mu_j}^T\Sigma^{-1}\underline{\mu_j} - 2\underline{\mu_i}^T\Sigma^{-1}\underline{X} + \underline{\mu_i}^T\Sigma^{-1}\underline{\mu_i}$$

$$2log\left[\frac{P(C_i)}{P(C_j)}\right] = 2\left[\underline{\mu_j}^T\Sigma^{-1} - \underline{\mu_i}^T\Sigma^{-1}\right]\underline{X} - \left[\underline{\mu_j}^T\Sigma^{-1}\underline{\mu_j} - \underline{\mu_i}^T\Sigma^{-1}\underline{\mu_i}\right]$$

Linear classifier!

Decision boundaries

- $2log\left[\frac{P(C_i)}{P(C_j)}\right] = 2\left[\underline{\mu}_j{}^T\Sigma^{-1} - \underline{\mu}_i{}^T\Sigma^{-1}\right]\underline{X} - \left[\underline{\mu}_j{}^T\Sigma^{-1}\underline{\mu}_j - \underline{\mu}_i{}^T\Sigma^{-1}\underline{\mu}_i\right]$

- Let:

$$\underline{W} = 2\left[\underline{\mu}_j{}^T\Sigma^{-1} - \underline{\mu}_i{}^T\Sigma^{-1}\right]^T = 2\left[\Sigma^{-1}(\underline{\mu}_j - \underline{\mu}_i)\right]$$

$$W_0 = -\left[\underline{\mu}_j{}^T\Sigma^{-1}\underline{\mu}_j - \underline{\mu}_i{}^T\Sigma^{-1}\underline{\mu}_i\right] - 2log\left[\frac{P(C_i)}{P(C_j)}\right]$$

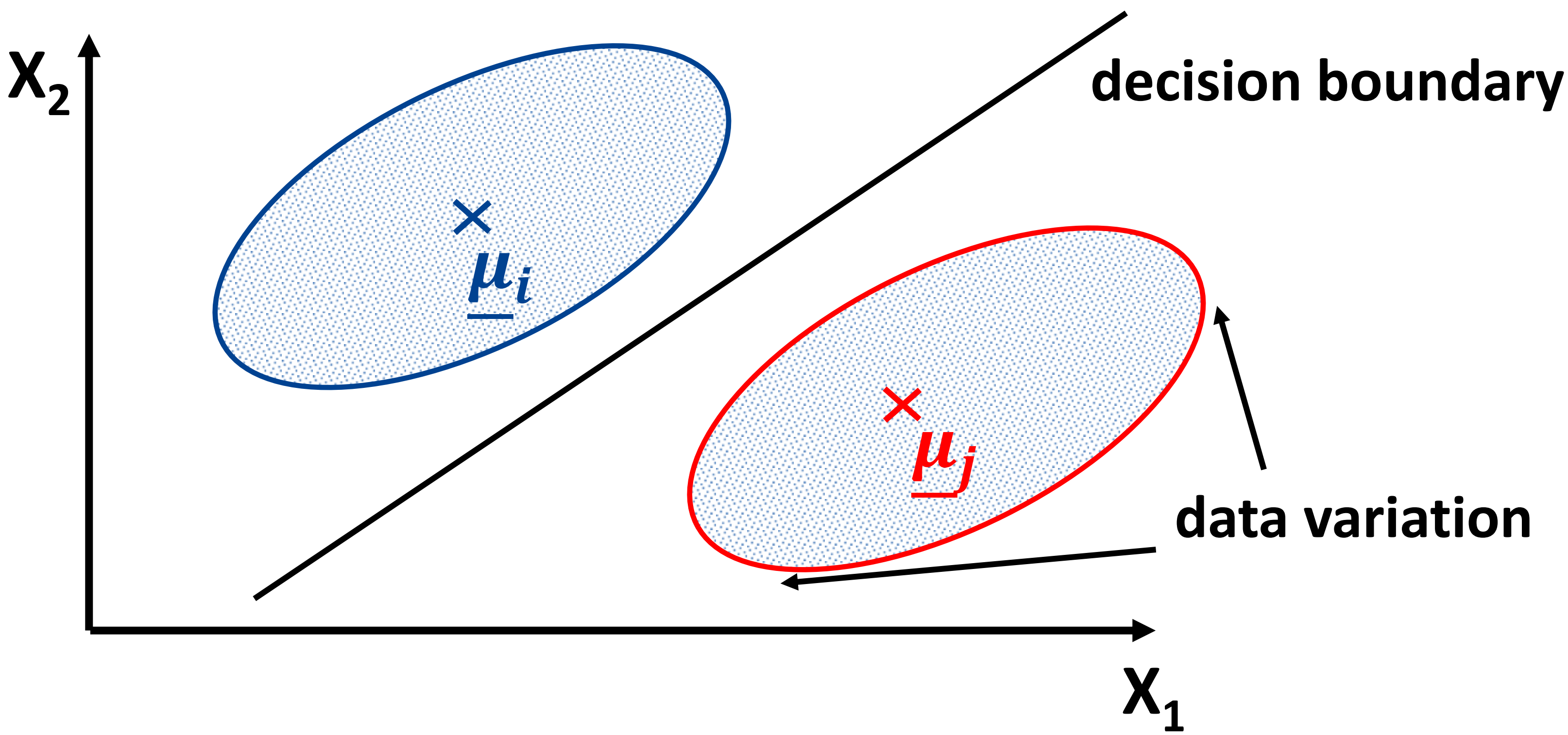
- Then decision boundary:

$\underline{W}^T\underline{X} + W_0 = 0$ → **Linear classifier!**

Decision boundaries

- Decision boundary:

  $\underline{W}^T \underline{X} + W_0 = 0$ → **Linear classifier!**



Scatter diagram

Applying Bayes Rule

- One way on how to apply Bayes rule in practical situations:
  - Obtain the training set $\underline{X}(1), \underline{X}(2) \cdots \underline{X}(M)$
  - Assume a multi-dimensional Gaussian density for each class, i.e., $P(\underline{X}|C_i)$
  - To obtain the form of each density we need $\underline{\mu_i}$ and Σ_i for each class i → estimate from training set
  - Estimate the a priori probabilities $P(C_i)$ from the training set, i.e., according to the frequencies of each class
  - Using the obtained estimates, plug in Bayes rule to obtain the classification rule

Estimate $\boldsymbol{\mu}$ and Σ

- Estimate $\boldsymbol{\mu}$ and Σ for a particular class:
  - We know that: $\underline{\mu} = E(\underline{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_N) \end{bmatrix}$
  - An estimate of $\mu_i \equiv E(X_i)$ is:

$$\hat{\mu}_i = \frac{1}{M} \sum_{m=1}^{M} X_i(m)$$

→ **the average**

$$\underline{\hat{\mu}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_N \end{bmatrix} = \frac{1}{M} \sum_{m=1}^{M} \underline{X}(m)$$

M is # of training patterns belonging to the considered class

Estimate μ and Σ

- Estimate $\boldsymbol{\mu}$ and Σ for a particular class:
  - We know that:
    estimate of variance $= \frac{1}{M}\sum_{m=1}^{M}(X_i(m) - \hat{\mu}_i)^2$
  - Est. of Σ : $\hat{\Sigma} = \frac{1}{M}\sum_{m=1}^{M}\left(\underline{X}(m) - \hat{\mu}\right)\left(\underline{X}(m) - \hat{\mu}\right)^T$
  - For terms on the diagonal, the previous formula reduces to: $\hat{\Sigma}_{(j,j)} = \frac{1}{M}\sum_{m=1}^{M}\left(X_j(m) - \hat{\mu}_j\right)^2$ which is equivalent to the estimate of variance formula
  - Confirming our earlier assertion that the terms on the diagonal of Σ give the variances of the components of $\underline{X}$

Acknowledgment

- These slides have been created relying on lecture notes of Prof. Dr. Amir Atiya