



# Recap

- We have studied the Bayes Classifier last class.



# Recap

- We have studied the Bayes Classifier last class.
- We have seen the Bayes classifier for general loss function and proved its optimality.

# Recap

- We have studied the Bayes Classifier last class.
- We have seen the Bayes classifier for general loss function and proved its optimality.
- We had also seen many special cases and how one can analytically derive the Bayes classifier for some simple cases of class conditional densities.

# An Example

Consider another example of deriving Bayes classifier.

- Suppose we have  $K$  classes. The classifier is allowed the option to 'reject' a pattern and this is done by the classifier assigning class  $K + 1$  to the pattern.

# An Example

Consider another example of deriving Bayes classifier.

- Suppose we have  $K$  classes. The classifier is allowed the option to 'reject' a pattern and this is done by the classifier assigning class  $K + 1$  to the pattern. Define the loss function by

$$\begin{aligned} L(i, j) &= 0 \text{ if } i = j \text{ and } i, j = 1, \dots, K \\ &= \rho_m \text{ if } i = 1, \dots, K, \text{ and } i \neq j \\ &= \rho_r \text{ if } i = K + 1 \end{aligned}$$

# An Example

Consider another example of deriving Bayes classifier.

- Suppose we have  $K$  classes. The classifier is allowed the option to 'reject' a pattern and this is done by the classifier assigning class  $K + 1$  to the pattern. Define the loss function by

$$\begin{aligned} L(i, j) &= 0 \text{ if } i = j \text{ and } i, j = 1, \dots, K \\ &= \rho_m \text{ if } i = 1, \dots, K, \text{ and } i \neq j \\ &= \rho_r \text{ if } i = K + 1 \end{aligned}$$

Now we want to derive the Bayes classifier in terms of the posterior probabilities.

## Example Contd.

- Recall that the Bayes classifier is

$$h_B(\mathbf{X}) = \alpha_i \quad \text{if}$$

$$R(\alpha_i | \mathbf{X}) \leq R(\alpha_j | \mathbf{X}), \quad \forall j.$$

where

$$R(\alpha_i | \mathbf{X}) = \sum_{j=0}^K L(\alpha_i, C_j) q_j(\mathbf{X})$$

## Example Contd.

- Recall that the Bayes classifier is

$$h_B(\mathbf{X}) = \alpha_i \quad \text{if}$$

$$R(\alpha_i | \mathbf{X}) \leq R(\alpha_j | \mathbf{X}), \quad \forall j.$$

where

$$R(\alpha_i | \mathbf{X}) = \sum_{j=0}^K L(\alpha_i, C_j) q_j(\mathbf{X})$$

- So, we now need to calculate  $R(\alpha_i | \mathbf{X})$  for different actions,  $\alpha_i$  available to the classifier.



- For  $\alpha_i = 1, \dots, K$ , we have  $L(\alpha_i, C_j) = \rho_m$  if  $\alpha_i \neq C_j$  and it is zero otherwise.

- For  $\alpha_i = 1, \dots, K$ , we have  $L(\alpha_i, C_j) = \rho_m$  if  $\alpha_i \neq C_j$  and it is zero otherwise.
- Hence,  $R(i | \mathbf{X}) = \sum_{j \neq i} \rho_m q_j(\mathbf{X}) = \rho_m(1 - q_i(\mathbf{X}))$ .

- For  $\alpha_i = 1, \dots, K$ , we have  $L(\alpha_i, C_j) = \rho_m$  if  $\alpha_i \neq C_j$  and it is zero otherwise.
- Hence,  $R(i | \mathbf{X}) = \sum_{j \neq i} \rho_m q_j(\mathbf{X}) = \rho_m(1 - q_i(\mathbf{X}))$ .
- Also,  $R(K + 1 | \mathbf{X}) = \sum_j \rho_r q_j(\mathbf{X}) = \rho_r$

- For  $\alpha_i = 1, \dots, K$ , we have  $L(\alpha_i, C_j) = \rho_m$  if  $\alpha_i \neq C_j$  and it is zero otherwise.
- Hence,  $R(i | \mathbf{X}) = \sum_{j \neq i} \rho_m q_j(\mathbf{X}) = \rho_m(1 - q_i(\mathbf{X}))$ .
- Also,  $R(K + 1 | \mathbf{X}) = \sum_j \rho_r q_j(\mathbf{X}) = \rho_r$
- Hence,  $h_B(\mathbf{X}) = i$ ,  $1 \leq i \leq K$ , if

- For  $\alpha_i = 1, \dots, K$ , we have  $L(\alpha_i, C_j) = \rho_m$  if  $\alpha_i \neq C_j$  and it is zero otherwise.
- Hence,  $R(i | \mathbf{X}) = \sum_{j \neq i} \rho_m q_j(\mathbf{X}) = \rho_m(1 - q_i(\mathbf{X}))$ .
- Also,  $R(K + 1 | \mathbf{X}) = \sum_j \rho_r q_j(\mathbf{X}) = \rho_r$
- Hence,  $h_B(\mathbf{X}) = i$ ,  $1 \leq i \leq K$ , if

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_m(1 - q_j(\mathbf{X})), \forall j$$

and

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_r$$

- We have,  $h_B(\mathbf{X}) = i$ ,  $1 \leq i \leq K$ , if

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_m(1 - q_j(\mathbf{X})), \forall j$$

and

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_r$$

- We have,  $h_B(\mathbf{X}) = i, 1 \leq i \leq K$ , if

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_m(1 - q_j(\mathbf{X})), \forall j$$

and

$$\rho_m(1 - q_i(\mathbf{X})) \leq \rho_r$$

- Thus,  $h_B(\mathbf{X}) = i, 1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .

• We saw,  $h_B(\mathbf{X}) = i, 1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .



- We saw,  $h_B(\mathbf{X}) = i$ ,  $1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X})$ ,  $\forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .

- If  $\rho_r \geq \rho_m$

- We saw,  $h_B(\mathbf{X}) = i$ ,  $1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X})$ ,  $\forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .

- If  $\rho_r \geq \rho_m$  – Never reject a pattern!

- We saw,  $h_B(\mathbf{X}) = i, 1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .

- If  $\rho_r \geq \rho_m$  – Never reject a pattern!
- If  $\rho_r = 0$

- We saw,  $h_B(\mathbf{X}) = i, 1 \leq i \leq K$ , if

(i).  $q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \forall j$ , and

(ii).  $q_i(\mathbf{X}) \geq 1 - \frac{\rho_r}{\rho_m}$ ;

else  $h_B(\mathbf{X}) = K + 1$ .

- If  $\rho_r \geq \rho_m$  – Never reject a pattern!
- If  $\rho_r = 0$  – Always reject the pattern (unless you are absolutely sure)



# Finding Bayes Error

- Given class conditional densities, the Bayes classifier is easily computed.

# Finding Bayes Error

- Given class conditional densities, the Bayes classifier is easily computed.
- We may also want to compute the Bayes error.
- Gives us the expected performance. Also lets us decide whether we need better features.

# Finding Bayes Error

- Given class conditional densities, the Bayes classifier is easily computed.
- We may also want to compute the Bayes error.
- Gives us the expected performance. Also lets us decide whether we need better features.
- For the case of 0-1 loss function, we need to evaluate

$$\int_{\mathcal{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

# Finding Bayes Error

- Given class conditional densities, the Bayes classifier is easily computed.
- We may also want to compute the Bayes error.
- Gives us the expected performance. Also lets us decide whether we need better features.
- For the case of 0-1 loss function, we need to evaluate

$$\int_{\mathcal{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

- In general, a difficult integral to evaluate.



- Let us consider the simplest case:  
2-class problem,  $X \in \mathfrak{R}$ , normal class conditional densities and 0-1 loss function.
- Assume equal priors. Let  $\sigma_0 = \sigma_1 = \sigma$  and  $\mu_0 < \mu_1$ .

- Let us consider the simplest case:  
2-class problem,  $X \in \mathfrak{R}$ , normal class conditional densities and 0-1 loss function.
- Assume equal priors. Let  $\sigma_0 = \sigma_1 = \sigma$  and  $\mu_0 < \mu_1$ .
- Then  $h_B(X) = 0$  if  $X < (\mu_0 + \mu_1)/2$ .

- Let us consider the simplest case:  
2-class problem,  $X \in \mathfrak{R}$ , normal class conditional densities and 0-1 loss function.
- Assume equal priors. Let  $\sigma_0 = \sigma_1 = \sigma$  and  $\mu_0 < \mu_1$ .
- Then  $h_B(X) = 0$  if  $X < (\mu_0 + \mu_1)/2$ .
- Then, Bayes error is

$$P(\text{error}) = 0.5 \int_{-\infty}^{\frac{\mu_0 + \mu_1}{2}} f_1(X) dX + 0.5 \int_{\frac{\mu_0 + \mu_1}{2}}^{\infty} f_0(X) dX$$

$$P(\text{error}) = 0.5 \int_{-\infty}^{\frac{\mu_0 + \mu_1}{2}} f_1(X) dX + 0.5 \int_{\frac{\mu_0 + \mu_1}{2}}^{\infty} f_0(X) dX$$

$$P(\text{error}) = 0.5 \int_{-\infty}^{\frac{\mu_0 + \mu_1}{2}} f_1(X) dX + 0.5 \int_{\frac{\mu_0 + \mu_1}{2}}^{\infty} f_0(X) dX$$

- Put  $Z = (X - \mu_1)/\sigma$  in the first and  $Z = (X - \mu_0)/\sigma$  in the second integral.

$$P(\text{error}) = 0.5 \int_{-\infty}^{\frac{\mu_0 + \mu_1}{2}} f_1(X) dX + 0.5 \int_{\frac{\mu_0 + \mu_1}{2}}^{\infty} f_0(X) dX$$

- Put  $Z = (X - \mu_1)/\sigma$  in the first and  $Z = (X - \mu_0)/\sigma$  in the second integral.
- Now both  $f_1$  and  $f_0$  become standard normal distribution.

$$P(\text{error}) = 0.5 \int_{-\infty}^{\frac{\mu_0 + \mu_1}{2}} f_1(X) dX + 0.5 \int_{\frac{\mu_0 + \mu_1}{2}}^{\infty} f_0(X) dX$$

- Put  $Z = (X - \mu_1)/\sigma$  in the first and  $Z = (X - \mu_0)/\sigma$  in the second integral.
- Now both  $f_1$  and  $f_0$  become standard normal distribution.
- The upper limit in the first integral becomes  $(\mu_0 - \mu_1)/2\sigma$  and lower limit in second integral becomes  $(\mu_1 - \mu_0)/2\sigma$ .

Now we get

$$P(\text{error}) = 0.5\Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) + 0.5\left[1 - \Phi\left(\frac{\mu_1 - \mu_0}{2\sigma}\right)\right]$$



Now we get

$$\begin{aligned} P(\text{error}) &= 0.5\Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) + 0.5\left[1 - \Phi\left(\frac{\mu_1 - \mu_0}{2\sigma}\right)\right] \\ &= \Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) \end{aligned}$$

Here,  $\Phi$  is the distribution function of the Standard Normal random Variable.

Now we get

$$\begin{aligned} P(\text{error}) &= 0.5\Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) + 0.5\left[1 - \Phi\left(\frac{\mu_1 - \mu_0}{2\sigma}\right)\right] \\ &= \Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) \end{aligned}$$

Here,  $\Phi$  is the distribution function of the Standard Normal random Variable.

The quantity  $\frac{|\mu_0 - \mu_1|}{\sigma}$  is called *discriminability*.

- In the general case, we need to evaluate

$$P(\text{error}) = \int_{\mathfrak{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

- In the general case, we need to evaluate

$$P(\text{error}) = \int_{\mathfrak{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

- A useful inequality here is

$$\min(a, b) \leq a^\beta b^{1-\beta}, \quad \forall a, b \geq 0, \quad 0 \leq \beta \leq 1.$$

- In the general case, we need to evaluate

$$P(\mathbf{error}) = \int_{\mathfrak{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

- A useful inequality here is

$$\min(a, b) \leq a^\beta b^{1-\beta}, \quad \forall a, b \geq 0, \quad 0 \leq \beta \leq 1.$$

- Easy to prove. Suppose  $a < b$

$$a^\beta b^{1-\beta} = a^{-1+\beta} b^{1-\beta} a = \left(\frac{b}{a}\right)^{1-\beta} a \geq a = \min(a, b)$$

- In the general case, we need to evaluate

$$P(\text{error}) = \int_{\mathfrak{R}^n} \min(p_0 f_0(\mathbf{X}), p_1 f_1(\mathbf{X})) d\mathbf{X}$$

- A useful inequality here is

$$\min(a, b) \leq a^\beta b^{1-\beta}, \quad \forall a, b \geq 0, \quad 0 \leq \beta \leq 1.$$


- Easy to prove. Suppose  $a < b$

$$a^\beta b^{1-\beta} = a^{-1+\beta} b^{1-\beta} a = \left(\frac{b}{a}\right)^{1-\beta} a \geq a = \min(a, b)$$

- Hence we have (for 0-1 loss function)

$$P(\text{error}) \leq p_0^\beta p_1^{1-\beta} \int_{\mathfrak{R}^n} f_0^\beta(\mathbf{X}) f_1^{1-\beta}(\mathbf{X}) d\mathbf{X}$$

•  
•  
•



Suppose  $f_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  and  $f_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ .

Suppose  $f_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$  and  $f_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ . Then we can show

$$\int f_0^\beta(\mathbf{X}) f_1^{1-\beta}(\mathbf{X}) d\mathbf{X} = \exp(-K(\beta))$$

where

$$K(\beta) = \frac{\beta(1-\beta)}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^t (\beta \Sigma_0 + (1-\beta) \Sigma_1)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \frac{1}{2} \ln \left( \frac{|\beta \Sigma_0 + (1-\beta) \Sigma_1|}{|\Sigma_0|^\beta |\Sigma_1|^{(1-\beta)}} \right)$$



- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$

- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$
- We can choose a  $\beta$  and calculate a bound from this expression.

- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$
- We can choose a  $\beta$  and calculate a bound from this expression.
- To get a tighter bound we can choose  $\beta$  to minimize  $\exp(-K(\beta))$ . Gives so called Chernoff bound.

- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$
- We can choose a  $\beta$  and calculate a bound from this expression.
- To get a tighter bound we can choose  $\beta$  to minimize  $\exp(-K(\beta))$ . Gives so called Chernoff bound.
- Often this minimization can be difficult.

- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$
- We can choose a  $\beta$  and calculate a bound from this expression.
- To get a tighter bound we can choose  $\beta$  to minimize  $\exp(-K(\beta))$ . Gives so called Chernoff bound.
- Often this minimization can be difficult.
- In such cases, a useful choice is  $\beta = 0.5$ . Known as Bhattacharya bound.

- We thus have:  $P(\text{error}) \leq p_0^\beta p_1^{(1-\beta)} \exp(-K(\beta))$
- We can choose a  $\beta$  and calculate a bound from this expression.
- To get a tighter bound we can choose  $\beta$  to minimize  $\exp(-K(\beta))$ . Gives so called Chernoff bound.
- Often this minimization can be difficult.
- In such cases, a useful choice is  $\beta = 0.5$ . Known as Bhattacharya bound.
- The bound  $\min(a, b) \leq a^\beta b^{(1-\beta)}$  can always be used; the resulting integral may be complex for other densities. Can use some numerical approximation.



# Other Criteria

- The Bayes classifier is optimal for the criterion of risk minimization.
- There can be other criteria.

# Other Criteria

- The Bayes classifier is optimal for the criterion of risk minimization.
- There can be other criteria.
- The Bayes classifier depends on both  $p_i$ , prior probabilities, and  $f_i$ , class conditional densities.
- Suppose we do not want to rely on prior probabilities.
- We may want a classifier that does best against any (or worst) prior probabilities.



- Consider a 2-class case.
- Let  $\mathcal{R}_i(h)$  denote the subset of feature space where  $h$  classifies into Class- $i$ .

- Consider a 2-class case.
- Let  $\mathcal{R}_i(h)$  denote the subset of feature space where  $h$  classifies into Class- $i$ .
- Then the Risk integral is

$$R(h) = \int_{\mathcal{R}_1(h)} L(1, 0)p_0f_0(\mathbf{X})d\mathbf{X} + \int_{\mathcal{R}_0(h)} L(0, 1)p_1f_1(\mathbf{X})d\mathbf{X}$$

- We can simplify this to get rid of dependence on priors.

- Using  $p_0 = 1 - p_1$ , we get

$$R = \int_{\mathcal{R}_1} L(1, 0)p_0 f_0(\mathbf{X}) d\mathbf{X} + \int_{\mathcal{R}_0} L(0, 1)p_1 f_1(\mathbf{X}) d\mathbf{X}$$

- Using  $p_0 = 1 - p_1$ , we get

$$\begin{aligned} R &= \int_{\mathcal{R}_1} L(1, 0)p_0 f_0(\mathbf{X}) d\mathbf{X} + \int_{\mathcal{R}_0} L(0, 1)p_1 f_1(\mathbf{X}) d\mathbf{X} \\ &= L(1, 0)p_0 \int_{\mathcal{R}_1} f_0(\mathbf{X}) d\mathbf{X} + \\ &\quad L(0, 1)(1 - p_0) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X} \end{aligned}$$

- Thus we get

$$\begin{aligned} R &= \int_{\mathcal{R}_1} L(1, 0) p_0 f_0(\mathbf{X}) d\mathbf{X} + \int_{\mathcal{R}_0} L(0, 1) p_1 f_1(\mathbf{X}) d\mathbf{X} \\ &= L(0, 1) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X} + \\ &\quad p_0 \left[ L(1, 0) \int_{\mathcal{R}_1} f_0(\mathbf{X}) d\mathbf{X} - L(0, 1) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X} \right] \end{aligned}$$

# Minmax Classifier

- Consider a classifier such that

$$L(1, 0) \int_{\mathcal{R}_1} f_0(\mathbf{X}) d\mathbf{X} = L(0, 1) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X}$$

# Minmax Classifier

- Consider a classifier such that

$$L(1, 0) \int_{\mathcal{R}_1} f_0(\mathbf{X}) d\mathbf{X} = L(0, 1) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X}$$

- For this classifier the risk would be independent of priors.

# Minmax Classifier

- Consider a classifier such that

$$L(1, 0) \int_{\mathcal{R}_1} f_0(\mathbf{X}) d\mathbf{X} = L(0, 1) \int_{\mathcal{R}_0} f_1(\mathbf{X}) d\mathbf{X}$$

- For this classifier the risk would be independent of priors.
- Called the minmax classifier
- We are minimizing the maximum possible (over all priors) risk.
- In general, finding the minmax classifier can be analytically complicated.





# Neyman-Pearson Criterion

- Bayes classifier minimizes risk.
- It minimizes some weighted sum of all errors.



# Neyman-Pearson Criterion

- Bayes classifier minimizes risk.
- It minimizes some weighted sum of all errors.
- We may not explicitly want to trade one type of error with another

# Neyman-Pearson Criterion

- Bayes classifier minimizes risk.
- It minimizes some weighted sum of all errors.
- We may not explicitly want to trade one type of error with another
- One criterion: minimize Type-II error under the constraint that Type-I error is below some threshold.
- This is the Neyman-Pearson criterion.

# Neyman-Pearson Criterion

- Bayes classifier minimizes risk.
- It minimizes some weighted sum of all errors.
- We may not explicitly want to trade one type of error with another
- One criterion: minimize Type-II error under the constraint that Type-I error is below some threshold.
- This is the Neyman-Pearson criterion.
- This could be useful in, e.g., biometric applications.

- Type-I error: Wrongly classifying a Class-0 pattern
- Suppose the upper bound on Type-I error is  $\alpha$ .
- The Neyman Person classifier can also be expressed as a threshold on the likelihood ratio.



# Neyman-Pearson Classifier

- The Neyman-Pearson classifier,  $h_{NP}$ , is characterized by: given any  $\alpha \in (0, 1)$

# Neyman-Pearson Classifier

- The Neyman-Pearson classifier,  $h_{NP}$ , is characterized by: given any  $\alpha \in (0, 1)$ 
  1.  $P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}] \leq \alpha$

# Neyman-Pearson Classifier

- The Neyman-Pearson classifier,  $h_{NP}$ , is characterized by: given any  $\alpha \in (0, 1)$

1.  $P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] \leq \alpha$

2.  $P[h_{NP}(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}] \leq [P[h(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}]$

for all  $h$  such that  $P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] \leq \alpha$



# Neyman-Person Classifier

- Let the bound on Type-I error be  $\alpha$ . Then

$$\begin{aligned}h_{NP}(\mathbf{X}) &= 1 \text{ if } \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > K \\ &= 0 \text{ Otherwise}\end{aligned}$$

where  $K$  is such that

$$P \left[ \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} \leq K \mid \mathbf{X} \in \mathbf{C-0} \right] = 1 - \alpha$$

(We assume  $P\{\mathbf{X} : f_1(\mathbf{X}) = K f_0(\mathbf{X})\} = 0$ , for simplicity)

- We now prove that this satisfies the NP Criterion. By construction, we have

$$\begin{aligned} P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] &= P \left[ \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > K \mid \mathbf{X} \in \mathbf{C-0} \right] \\ &= \alpha \end{aligned}$$

- We now prove that this satisfies the NP Criterion. By construction, we have

$$\begin{aligned} P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] &= P \left[ \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > K \mid \mathbf{X} \in \mathbf{C-0} \right] \\ &= \alpha \end{aligned}$$

- So, we need to show that its Type-II error is less than that for any other classifier satisfying the constraint on Type-I error.

- Let  $h$  be any classifier such that

$$P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] \leq \alpha$$

- Let  $h$  be any classifier such that

$$P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] \leq \alpha$$

- To complete the proof we have to show that

$$P[h_{NP}(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}] \leq P[h(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}]$$

- Let  $h$  be any classifier such that

$$P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] \leq \alpha$$

- To complete the proof we have to show that

$$P[h_{NP}(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}] \leq P[h(\mathbf{X}) = 0 \mid \mathbf{X} \in \mathbf{C-1}]$$

Or, equivalently

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}] \geq [P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}]$$

- Consider the Integral

$$I = \int_{\mathbb{R}^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x})) (f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x}$$

- Consider the Integral

$$\begin{aligned} I &= \int_{\mathbb{R}^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x})) (f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x} \\ &= \int_{f_1 > K f_0} (h_{NP}(\mathbf{x}) - h(\mathbf{x})) (f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x} + \\ &\quad \int_{f_1 \leq K f_0} (h_{NP}(\mathbf{x}) - h(\mathbf{x})) (f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x} \end{aligned}$$

- We first show that this integral is always non-negative.



- When  $f_1(\mathbf{x}) > K f_0(\mathbf{x})$ , we have  
$$h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 1 - h(\mathbf{x}) \geq 0$$

- When  $f_1(\mathbf{x}) > K f_0(\mathbf{x})$ , we have  
 $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 1 - h(\mathbf{x}) \geq 0$  which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) \geq 0$$

- When  $f_1(\mathbf{x}) > K f_0(\mathbf{x})$ , we have  
 $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 1 - h(\mathbf{x}) \geq 0$  which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) \geq 0$$

- Similarly, when  $f_1(\mathbf{x}) < K f_0(\mathbf{x})$ , we have  
 $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 0 - h(\mathbf{x}) \leq 0$  which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) \geq 0$$

- When  $f_1(\mathbf{x}) > K f_0(\mathbf{x})$ , we have  
 $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 1 - h(\mathbf{x}) \geq 0$  which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) \geq 0$$

- Similarly, when  $f_1(\mathbf{x}) < K f_0(\mathbf{x})$ , we have  
 $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 0 - h(\mathbf{x}) \leq 0$  which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) \geq 0$$

- This shows that  $I \geq 0$ .

- Thus, we have

$$\int_{\mathbb{R}^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x} \geq 0$$

- Thus, we have

$$\int_{\mathbb{R}^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - K f_0(\mathbf{x})) d\mathbf{x} \geq 0$$

- This implies

$$\int h_{NP}(\mathbf{x}) f_1(\mathbf{x}) d\mathbf{x} - \int h(\mathbf{x}) f_1(\mathbf{x}) d\mathbf{x} \geq$$

$$K \left[ \int h_{NP}(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} - \int h(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \right]$$

•  
•  
•

Since  $h_{NP}$  and  $h$  take values in  $\{0, 1\}$ ,

$$\int_{\mathbb{R}^n} h_{NP}(\mathbf{x}) f_1(\mathbf{X}) d\mathbf{X} = P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}]$$

and

$$\int_{\mathbb{R}^n} h(\mathbf{x}) f_1(\mathbf{X}) d\mathbf{X} = P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}]$$

•  
•  
•



Since  $h_{NP}$  and  $h$  take values in  $\{0, 1\}$ ,

$$\int_{\mathbb{R}^n} h_{NP}(\mathbf{x}) f_1(\mathbf{X}) d\mathbf{X} = P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}]$$

and

$$\int_{\mathbb{R}^n} h(\mathbf{x}) f_1(\mathbf{X}) d\mathbf{X} = P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}]$$

Similarly for the integrals involving  $f_0$ .



- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] \geq$$

$$K [P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}]]$$

- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-1}] \geq$$

$$K [P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C-0}]]$$

- But for all  $h$  under consideration, the RHS above is non-negative.

- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] \geq$$

$$K [P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}]]$$

- But for all  $h$  under consideration, the RHS above is non-negative. Hence

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] \geq 0$$

- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] \geq$$

$$K [P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-0}]]$$

- But for all  $h$  under consideration, the RHS above is non-negative. Hence

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{c-1}] \geq 0$$

- This completes the proof.

- Neymann-Pearson classifier also needs knowledge of class conditional densities.
- Like Bayes classifier, it also is based on the ratio  $\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}$ .
- In Bayes classifier we say **c-1** if  $\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > \frac{p_0}{p_1} \frac{L(0,1)}{L(1,0)}$ .
- In NP, this threshold,  $K$ , is set based on the allowed Type-I error.

# Example of NP classifier

- Take  $X \in \mathcal{R}$  and class conditional densities normal with equal variance. Let  $\mu_0 < \mu_1$ .
- Now the NP classifier is: If  $X > \tau$  then **c-1** where  $\tau$  is simply determined by Type-I error bound.
- This is intuitively clear.
- We will now derive this formally.

# Example

Now (assuming  $\mu_1 > \mu_0$ ),

$$\frac{f_1(X)}{f_0(X)} = \exp\left(-\frac{(X - \mu_1)^2}{2\sigma^2} + \frac{(X - \mu_0)^2}{2\sigma^2}\right)$$

# Example

Now (assuming  $\mu_1 > \mu_0$ ),

$$\begin{aligned}\frac{f_1(X)}{f_0(X)} &= \exp\left(-\frac{(X - \mu_1)^2}{2\sigma^2} + \frac{(X - \mu_0)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}[\mu_1^2 - \mu_0^2 - 2X(\mu_1 - \mu_0)]\right)\end{aligned}$$



# Example

Now (assuming  $\mu_1 > \mu_0$ ),

$$\begin{aligned}\frac{f_1(X)}{f_0(X)} &= \exp\left(-\frac{(X - \mu_1)^2}{2\sigma^2} + \frac{(X - \mu_0)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}[\mu_1^2 - \mu_0^2 - 2X(\mu_1 - \mu_0)]\right) \\ &= \exp\left(\frac{\mu_1 - \mu_0}{2\sigma^2}[2X - (\mu_1 + \mu_0)]\right)\end{aligned}$$

- We need to find  $K$  such that

$$P \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \mid X \in \mathbf{C-0} \right] = 1 - \alpha$$

- We need to find  $K$  such that

$$P \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \mid X \in \mathbf{C-0} \right] = 1 - \alpha$$

- From the earlier expression,  $\ln \frac{f_1(X)}{f_0(X)} \leq \ln K$  is same as

$$\frac{\mu_1 - \mu_0}{2\sigma^2} [2X - (\mu_1 + \mu_0)] \leq \ln K$$

- We need to find  $K$  such that

$$P \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \mid X \in \mathbf{C-0} \right] = 1 - \alpha$$

- From the earlier expression,  $\ln \frac{f_1(X)}{f_0(X)} \leq \ln K$  is same as

$$\frac{\mu_1 - \mu_0}{2\sigma^2} [2X - (\mu_1 + \mu_0)] \leq \ln K$$

$$i.e., \quad X \leq \frac{\sigma^2 \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2}$$

Hence we have ( writing  $P[A|X \in \mathbf{c-0}]$  as  $P_0[A]$ )

Hence we have (writing  $P[A|X \in \mathbf{c-0}]$  as  $P_0[A]$ )

$$P_0 \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \right] = P_0 \left[ X \leq \frac{\sigma^2 \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2} \right]$$

Hence we have (writing  $P[A|X \in \mathbf{c-0}]$  as  $P_0[A]$ )

$$\begin{aligned} P_0 \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \right] &= P_0 \left[ X \leq \frac{\sigma^2 \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2} \right] \\ &= P_0 \left[ \frac{X - \mu_0}{\sigma} \leq \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right] \end{aligned}$$

Hence we have (writing  $P[A|X \in \mathbf{c-0}]$  as  $P_0[A]$ )

$$\begin{aligned} P_0 \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \right] &= P_0 \left[ X \leq \frac{\sigma^2 \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2} \right] \\ &= P_0 \left[ \frac{X - \mu_0}{\sigma} \leq \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right] \\ &= \Phi \left( \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right) \end{aligned}$$



Hence we have (writing  $P[A|X \in \mathbf{c-0}]$  as  $P_0[A]$ )

$$\begin{aligned} P_0 \left[ \ln \frac{f_1(X)}{f_0(X)} \leq \ln K \right] &= P_0 \left[ X \leq \frac{\sigma^2 \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2} \right] \\ &= P_0 \left[ \frac{X - \mu_0}{\sigma} \leq \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right] \\ &= \Phi \left( \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right) \end{aligned}$$

We need this quantity to be equal to  $(1 - \alpha)$ .

Thus we want

$$\Phi \left( \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right) = (1 - \alpha)$$

Thus we want

$$\Phi \left( \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right) = (1 - \alpha)$$

This gives us an expression for  $\ln K$  as

$$\frac{\sigma \ln K}{\mu_1 - \mu_0} = \Phi^{-1}(1 - \alpha) - \frac{\mu_1 - \mu_0}{2\sigma}$$

Thus we want

$$\Phi \left( \frac{\sigma \ln K}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma} \right) = (1 - \alpha)$$


This gives us an expression for  $\ln K$  as

$$\frac{\sigma \ln K}{\mu_1 - \mu_0} = \Phi^{-1}(1 - \alpha) - \frac{\mu_1 - \mu_0}{2\sigma}$$

or

$$\ln K = \frac{\mu_1 - \mu_0}{\sigma} \Phi^{-1}(1 - \alpha) - \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$$

•  
•  
•



We say  $X \in \mathbf{c-1}$  if  $\ln \frac{f_1(X)}{f_0(X)} > \ln K$ . That is

•  
•  
•

We say  $X \in \mathbf{c-1}$  if  $\ln \frac{f_1(X)}{f_0(X)} > \ln K$ . That is

$$\frac{\mu_1 - \mu_0}{2\sigma^2} [2X - (\mu_1 + \mu_0)] > \frac{\mu_1 - \mu_0}{\sigma} \Phi^{-1}(1 - \alpha) - \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$$

•  
•  
•

We say  $X \in \mathbf{c-1}$  if  $\ln \frac{f_1(X)}{f_0(X)} > \ln K$ . That is

$$\frac{\mu_1 - \mu_0}{2\sigma^2} [2X - (\mu_1 + \mu_0)] > \frac{\mu_1 - \mu_0}{\sigma} \Phi^{-1}(1 - \alpha) - \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$$

*i.e.*,  $2X - (\mu_1 + \mu_0) > 2\sigma \Phi^{-1}(1 - \alpha) - (\mu_1 - \mu_0)$

•  
•  
•

We say  $X \in \mathbf{c-1}$  if  $\ln \frac{f_1(X)}{f_0(X)} > \ln K$ . That is

$$\frac{\mu_1 - \mu_0}{2\sigma^2} [2X - (\mu_1 + \mu_0)] > \frac{\mu_1 - \mu_0}{\sigma} \Phi^{-1}(1 - \alpha) - \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$$

*i.e.*,  $2X - (\mu_1 + \mu_0) > 2\sigma \Phi^{-1}(1 - \alpha) - (\mu_1 - \mu_0)$

*i.e.*,  $X > \sigma \Phi^{-1}(1 - \alpha) + \mu_0$



•  
•  
•



Thus NP classifier puts  $X \in \mathbf{c-1}$  if

$$i.e., \quad X > \sigma \Phi^{-1}(1 - \alpha) + \mu_0$$

Thus NP classifier puts  $X \in \mathbf{c-1}$  if

$$i.e., \quad X > \sigma \Phi^{-1}(1 - \alpha) + \mu_0$$

$$i.e., \quad \Phi\left(\frac{X - \mu_0}{\sigma}\right) > (1 - \alpha)$$

Thus NP classifier puts  $X \in \mathbf{c-1}$  if

$$i.e., \quad X > \sigma \Phi^{-1}(1 - \alpha) + \mu_0$$

$$i.e., \quad \Phi\left(\frac{X - \mu_0}{\sigma}\right) > (1 - \alpha)$$

This means the NP classifier puts  $X$  in  $\mathbf{c-1}$  if  $X > \tau$

where  $\int_{\tau}^{\infty} f_0(X) dX = \alpha$ .

- Like the Bayes classifier, the NP classifier also needs knowledge of class conditional densities.
- NP classifier is only for the 2-class case.
- It is actually more important in **hypothesis testing** problems. (Likelihood ratio test)

# Receiver Operating Characteristic (ROC)

- Consider a one dimensional feature space, 2-class problem with a classifier,  $h(X) = 0$  if  $X < \tau$ .
- Consider equal priors, Gaussian class conditional densities with equal variance, 0-1 loss. Now let us write the probability of error as a function of  $\tau$ .

# Receiver Operating Characteristic (ROC)

$$\begin{aligned} P[\text{error}] &= 0.5 \int_{-\infty}^{\tau} f_1(X) dX + 0.5 \int_{\tau}^{\infty} f_0(X) dX \\ &= 0.5 \Phi \left( \frac{\tau - \mu_1}{\sigma} \right) + 0.5 \left( 1 - \Phi \left( \frac{\tau - \mu_0}{\sigma} \right) \right) \end{aligned}$$

- As we vary  $\tau$  we trade one kind of error with another. In Bayes classifier, the loss function determines the 'exchange rate'.

# ROC curve

- The receiver operating characteristic (ROC) curve is one way to conveniently visualize and exploit this trade off.
- For a two class classifier there are four possible outcomes of a classification decision – two are correct decisions and two are errors.
- Let  $e_i$  denote probability of wrongly assigning class  $i$ ,  $i = 0, 1$ .

# ROC curve

Then we have

$$e_0 = P[X \leq \tau \mid X \in \mathbf{c-1}] \quad (\text{a miss})$$

$$e_1 = P[X > \tau \mid X \in \mathbf{c-0}] \quad (\text{false alarm})$$

$$1 - e_0 = P[X > \tau \mid X \in \mathbf{c-1}] \quad (\text{correct detection})$$

$$1 - e_1 = P[X \leq \tau \mid X \in \mathbf{c-0}] \quad (\text{correct rejection})$$



# ROC curve

Then we have

$$e_0 = P[X \leq \tau \mid X \in \mathbf{c-1}] \quad (\text{a miss})$$

$$e_1 = P[X > \tau \mid X \in \mathbf{c-0}] \quad (\text{false alarm})$$

$$1 - e_0 = P[X > \tau \mid X \in \mathbf{c-1}] \quad (\text{correct detection})$$

$$1 - e_1 = P[X \leq \tau \mid X \in \mathbf{c-0}] \quad (\text{correct rejection})$$

- For fixed class conditional densities, if we vary  $\tau$  the point  $(e_1, 1 - e_0)$  moves on a smooth curve in  $\mathfrak{R}^2$ .
- This is traditionally called the ROC curve. (Choice of coordinates is arbitrary)

- For any fixed  $\tau$  we can estimate  $e_0$  and  $e_1$  from training data.

- For any fixed  $\tau$  we can estimate  $e_0$  and  $e_1$  from training data.
- Hence, varying  $\tau$  we can find ROC and decide which may be the best operating point.

- For any fixed  $\tau$  we can estimate  $e_0$  and  $e_1$  from training data.
- Hence, varying  $\tau$  we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.

- For any fixed  $\tau$  we can estimate  $e_0$  and  $e_1$  from training data.
- Hence, varying  $\tau$  we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.
- When the class conditional densities are Gaussian with equal variance, we use this procedure to estimate Bayes error also.

- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$

$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$

$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then,  $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$ , the discriminability.

- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$

$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then,  $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$ , the discriminability.
- Knowing  $e_1$ ,  $(1 - e_0)$ , we can get  $d$  and hence the Bayes error. For our given  $\tau$  we can also get the actual error probability. We can tweak  $\tau$  to match the Bayes error.



- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$

$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then,  $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$ , the discriminability.
- Knowing  $e_1, (1 - e_0)$ , we can get  $d$  and hence the Bayes error. For our given  $\tau$  we can also get the actual error probability. We can tweak  $\tau$  to match the Bayes error.

- We can in general use the ROC curve in multidimensional cases also. Consider, for example,

$$h(\mathbf{X}) = \text{sgn}(\mathbf{W}^t \mathbf{X} + w_0).$$

We can use ROC to fix  $w_0$  after learning  $\mathbf{W}$ .



# Summary

- Bayes classifier is optimal for minimizing risk.



# Summary

- Bayes classifier is optimal for minimizing risk.
- we can derive Bayes classifier if we know class conditional densities.



# Summary

- Bayes classifier is optimal for minimizing risk.
- we can derive Bayes classifier if we know class conditional densities.
- There are criteria other than minimizing risk.

# Summary

- Bayes classifier is optimal for minimizing risk.
- we can derive Bayes classifier if we know class conditional densities.
- There are criteria other than minimizing risk.
- MinMax classifier, Neymann-Pearson Classifier are some such examples.

# Summary

- Bayes classifier is optimal for minimizing risk.
- we can derive Bayes classifier if we know class conditional densities.
- There are criteria other than minimizing risk.
- MinMax classifier, Neymann-Pearson Classifier are some such examples.
- ROC curves allow us to visualize trade-offs between different types of errors as we vary a threshold.

- 
- 
- 



- 
- 
- 
- 
- 
- 
- 
- 
- 
-