Lecture Notes on

Pattern Recognition

Session 02(B): Bayesian Decision Theory (II)

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

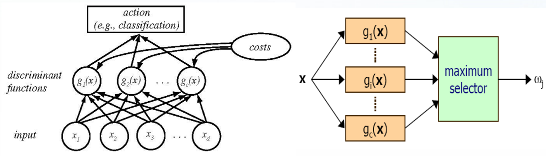
Road map

Discriminant Functions and Decision Surfaces

- Discriminants
- Decision Surfaces

Discriminants

- Many different ways to represent pattern classifiers; one of the most useful is in terms of discriminant functions



Decide ω_i if $g(x_i) > g(x_j);\ \forall j \neq i$

Discriminants for Bayes Classifier

- Using risks:

$$g_i(x) = -R(\alpha_i \mid x)$$

- Using zero-one loss function
  - i.e., min error rate

$$g_i(x) = P(\omega_i \mid x)$$

Discriminants for Bayes Classifier

- Is the choice of g_i unique?
  - Replacing $g_i(x)$ with $f(g_i(x))$, where f( ) is monotonically increasing, does not change the classification results.

$$g_i(\mathbf{x}) = p(\omega_i \mid x) \Rightarrow \begin{aligned} g_i(\mathbf{x}) &= \frac{p(\mathbf{x}/\omega_j)P(\omega_j)}{p(\mathbf{x})} \\ g_i(\mathbf{x}) &= p(\mathbf{x}/\omega_j)P(\omega_j) \\ g_i(\mathbf{x}) &= \ln p(\mathbf{x}/\omega_j) + \ln P(\omega_j) \end{aligned}$$

Discriminants for Two Categories

- A classifier is a "**dichotomizer**" that has two discriminant functions g_1 and g_2

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

Decide ω_1 if $g(x) > 0$; otherwise ω_2

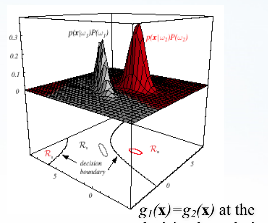
- Computation of $g(x) = g_1(x) - g_2(x)$

$$g(\mathbf{x}) = P(\omega_1 / \mathbf{x}) - P(\omega_2 / \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Decision Regions and Boundaries

- Effect of any decision rule is to divide the feature space into c decision regions, separated by decision boundaries.
  - Region R_i means assign x to ω_i



$g_1(\mathbf{x})=g_2(\mathbf{x})$ at the decision boundaries

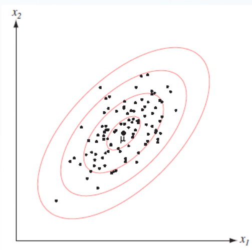
Decide ω_1 if $g_i(x) > g_j(x)$, $\forall j \neq i$ then $\boldsymbol{x}$ is in R_i

Discriminant Function for Multivariate Gaussian

- Multivariate Gaussian
- Discriminants for Gaussian
  - Case 1
  - Case 2
  - Case 3
  - Case 4

Multivariate Gaussian

- Samples drawn from a normal population tend to fall in a single cloud or cluster;
  - cluster center is determined by the mean vector and
  - shape by the covariance matrix
- The loci of points of constant density are hyper-ellipsoids whose principal axes are the eigenvectors of Σ



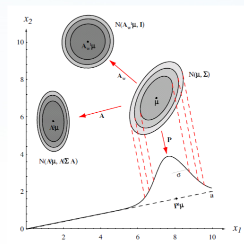
Multivariate Gaussian

$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \right]$$

- Multivariate density: N(μ , Σ) in d dimensions:
  - $x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose of a vector)
  - $\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector
  - Σ = d*d covariance matrix
  - $|\Sigma|$ and Σ^{-1} are determinant and inverse of Σ, respectively
- The covariance matrix is always symmetric and positive semi-definite;
  - we assume Σ is positive definite so its determinant is strictly positive
- Multivariate normal density is specified by d + d(d+1)/2 parameters
- If variables x_i and x_j are statistically independent then the covariance of x_i and x_j is zero.

Multivariate Gaussian

- Linear combinations of jointly normally distributed random variables are normally distributed
- Coordinate transformation can convert an arbitrary multivariate normal distribution into a spherical one



Discriminant Function for Multivariate Gaussian

- The minimum error-rate classification can be achieved by the discriminant function

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

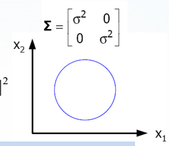
- In case of multivariate normal densities

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Discriminant Function for Multivariate Gaussian (Case 1)

- Case 1: $g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{1}{2}\ln|\Sigma_i|$
  - all priors are equal $p(\omega_1) = \ldots = p(\omega_c) = 1/c$
  - Case 1A: $\sum_i = \sigma^2 I$
    - Features are statistically independent
    - Each feature has the same variance

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^t(x-\mu_i) = -\frac{1}{2\sigma^2}\|x-\mu_i\|^2$$

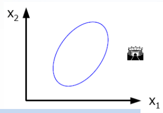


Minimum Distance Classifier: Euclidean Distance

Discriminant Function for Multivariate Gaussian (Case 1)

- Case 1: $g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{1}{2}\ln|\Sigma_i|$
  - all priors are equal $p(\omega_1) = \ldots = p(\omega_c) = 1/c$
  - Case 1B: $\sum_i = \sum$
    - Each category has the same covariance matrix

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i)$$



Minimum Distance Classifier: Mahalanobis Distance

Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\sum_i = \sigma^2 I$ $\quad g_i(x) = -\frac{1}{2}(x-\mu_i)^t(x-\mu_i) + \ln P(\omega_i)$
  - Features are statistically independent and
  - Each feature has the same variance

$g_i(x) = w_i^t x + w_{i0}$ (linear discriminant function)

where:

$$w_i = \frac{\mu_i}{\sigma^2}; \quad w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)$$

(ω_{i0} is called the threshold for the ***i***-th category!)

Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\sum_i = \sigma^2 I$
  - A classifier that uses linear discriminant functions is called "**a linear machine**"
  - Its decision surfaces are pieces of hyper-planes defined by the linear equations: $g_i(x) = g_j(x)$

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

$$\Leftrightarrow \left[2\boldsymbol{\mu}_i{}^t\mathbf{x} - 2\boldsymbol{\mu}_j{}^t\mathbf{x} - \boldsymbol{\mu}_i{}^t\boldsymbol{\mu}_i + \boldsymbol{\mu}_j{}^t\boldsymbol{\mu}_j\right] + 2\sigma^2 \log\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(2\mathbf{x} - \boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + 2\sigma^2 \log\frac{P(\omega_i)}{P(\omega_j)}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} = 0$$

$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\left[\mathbf{x} - \left(\underbrace{\boxed{\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\log\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}}_{\mathbf{x}_0}\right)\right] = 0$$

Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\sum_i = \sigma^2 I$
  - The hyper-plane separating R_i and R_j
    - passes through x_0
    - is orthogonal to the line linking the means.

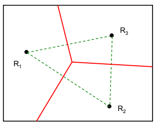
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\Sigma_i = \sigma^2 I$
  - Now consider the case 1A: equal priors
    - Discriminant function depends on the means only

$$g_i(x) = \frac{\mu_i^t}{\sigma^2} x - \frac{1}{2\sigma^2}\mu_i^t \mu_i \approx 2\mu_i^t x - \mu_i^t \mu_i$$

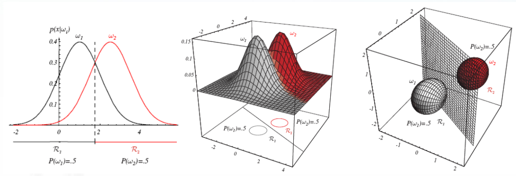


- Decision surface $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \left(\mathbf{x} - \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2}\right) = 0$

$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

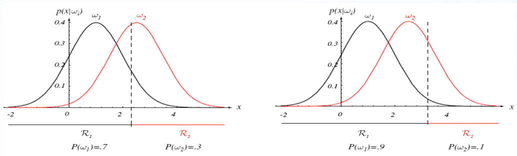
Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\Sigma_i = \sigma^2 I$
  - with equal priors



Discriminant Function for Multivariate Gaussian (Case 2)

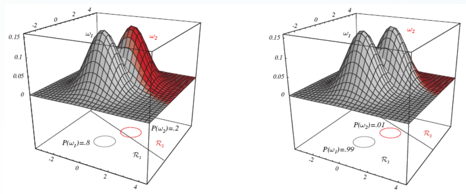
- Case 2: $\Sigma_i = \sigma^2 I$
  - with different priors



$P(\omega_1) = 0.7 \quad P(\omega_2) = 0.3$ $\qquad$ $P(\omega_1) = 0.9 \quad P(\omega_2) = 0.1$

Discriminant Function for Multivariate Gaussian (Case 2)

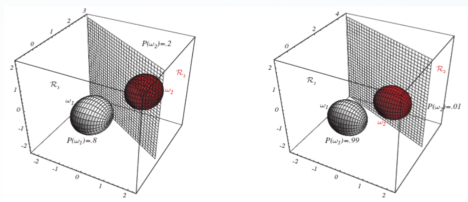
- Case 2: $\Sigma_i = \sigma^2 I$
  - with different priors



$P(\omega_1) = 0.8 \quad P(\omega_2) = 0.2$ $\qquad$ $P(\omega_1) = 0.99 \quad P(\omega_2) = 0.01$

Discriminant Function for Multivariate Gaussian (Case 2)

- Case 2: $\Sigma_i = \sigma^2 I$
  - with different priors



$P(\omega_1) = 0.8 \quad P(\omega_2) = 0.2$ $\qquad$ $P(\omega_1) = 0.99 \quad P(\omega_2) = 0.01$

Discriminant Function for Multivariate Gaussian (Case 3)

- Case 3: $\Sigma_i = \Sigma$
  - covariance matrices of all classes are identical but otherwise arbitrary!

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \log P(\omega_i)$$

Same for all i

$$= -\frac{1}{2}\left[\mathbf{x}^t \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \log P(\omega_i)$$

$$= -\frac{1}{2}\left[-2\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \log P(\omega_i)$$

Discriminant Function for Multivariate Gaussian (Case 3)

- Case 3: $\Sigma_i = \Sigma$
  - Hyper-plane separating R_i and R_j

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$
$$\Leftrightarrow \left[2\,\boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\,\boldsymbol{\mu}_j^t\,\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right] + 2\log\frac{P(\omega_i)}{P(\omega_j)} = 0$$
$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(2\mathbf{x} - \boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + 2\log\frac{P(\omega_i)}{P(\omega_j)}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} = 0$$
$$\Leftrightarrow \left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]^t\left[\mathbf{x} - \underbrace{\boxed{\left(\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\log\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right)}}_{\mathbf{x}_0}\right] = 0$$

The hyper-plane separating R_i and R_j is generally not orthogonal to the line between the means!

Discriminant Function for Multivariate Gaussian (Case 3)

- Case 3: $\Sigma_i = \Sigma$
  - Now consider the case 1B: equal priors
    - Discriminant function will be

$$g_i(\mathbf{x}) = -2\,\boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

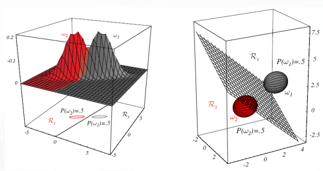
    - Decision Surface

$$\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]^t\left(\mathbf{x} - \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2}\right) = 0$$

- The hyper-plane passes through the center of the line between the means, though not orthogonal to it.

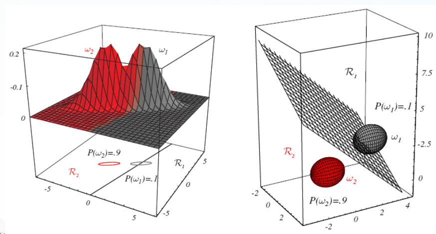
Discriminant Function for Multivariate Gaussian (Case 3)

- Case 3: $\Sigma_i = \Sigma$
  - Now consider the case 1B: equal priors
  - The hyper-plane passes through the center of the line between the means, though not orthogonal to it.



Discriminant Function for Multivariate Gaussian (Case 3)

- Case 3: $\Sigma_i = \Sigma$
  - If with different priors



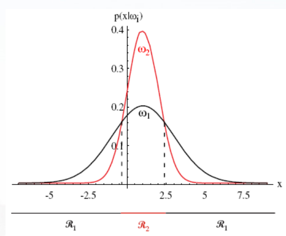
Discriminant Function for Multivariate Gaussian (Case 4)

- Case 4: Σ_i = arbitrary
  - The covariance matrices are different for each category

$$g_i(x)$$
$$= -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$
$$= x^t W_i x + w_i^t x + w_{i0}$$
$$where:$$
$$\mathrm{W}_\mathrm{i} = -\frac{1}{2}\Sigma_i^{-1}, \quad \mathrm{w}_\mathrm{i} = \Sigma_i^{-1}\mu_i$$
$$\mathrm{w}_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

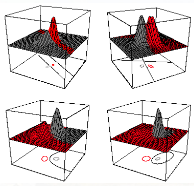
Discriminant Function for Multivariate Gaussian (Case 4)

- Case 4: Σ_i = arbitrary
  - Discriminant Functions for 1D Gaussian



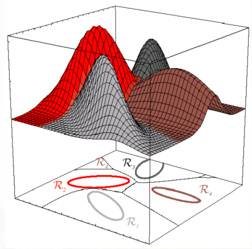
Discriminant Function for Multivariate Gaussian (Case 4)

- Case 4: Σ_i = arbitrary
  - In the 2-category case, decision surfaces are hyperquadrics that can assume any general form:
    - hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids



Discriminant Function for Multivariate Gaussian (Case 4)

- Case 4: Σ_i = arbitrary
  - More general cases



Discriminant Function for Multivariate Gaussian (Case 4)

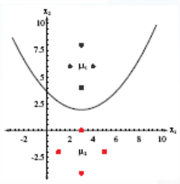
- Case 4: Σ_i = arbitrary
  - Example

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5,$$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$



Error Bounds

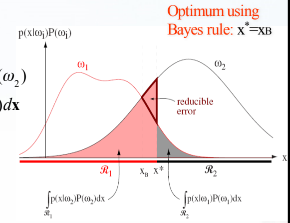
- Chernoff Bound
- Bhattacharyya Bound

Error Probabilities and Integrals

- Two types of errors for 2-class problem

$$P(error)$$
$$= p(\mathbf{x} \in R_2, \omega_1) + p(\mathbf{x} \in R_1, \omega_2)$$
$$= p(\mathbf{x} \in R_2 / \omega_1) P(\omega_1) + p(\mathbf{x} \in R_1 / \omega_2) P(\omega_2)$$
$$= \int_{R_2} p(\mathbf{x} / \omega_1) P(\omega_1) d\mathbf{x} + \int_{R_1} p(\mathbf{x} / \omega_2) P(\omega_2) d\mathbf{x}$$

Optimum using Bayes rule: $x^* = x_B$



Error Bounds

- Chernoff Bound

$$P(error) = \int P(error, \mathbf{x})\, d\mathbf{x} = \int P(error/\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}$$

$$P(error/\mathbf{x}) = \begin{cases} P(\omega_1/\mathbf{x}) & \text{if we decide } \omega_2 \\ P(\omega_2/\mathbf{x}) & \text{if we decide } \omega_1 \end{cases}$$

- Using the inequality:

$$min[a, b] \le a^\beta b^{1-\beta}, \qquad a, b \ge 0, 0 \le \beta \le 1$$

or $P(error/x) = \min[P(\omega_1/x), P(\omega_2/x)]$

$$P(error) = \int min[p(\mathbf{x}/\omega_1)P(\omega_1), p(\mathbf{x}/\omega_2)P(\omega_2)] d\mathbf{x} \le$$

$$P^\beta(\omega_1) P^{1-\beta}(\omega_2) \int p^\beta(\mathbf{x}/\omega_1)\, p^{1-\beta}(\mathbf{x}/\omega_2) d\mathbf{x}$$

Error Bounds

- Chernoff Bound
  - If the class conditional distributions are ***Gaussians***, then

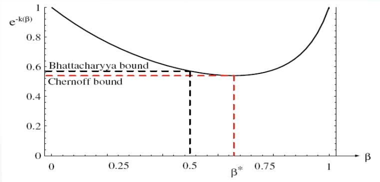
$$\int p^{\beta}(\mathbf{x}/\omega_1)\, p^{1-\beta}(\mathbf{x}/\omega_2) d\mathbf{x} = e^{-\kappa(\beta)}$$

where

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2-\mu_1)^t[\beta\Sigma_1+(1-\beta)\Sigma_2]^{-1}(\mu_2-\mu_1)+ \frac{1}{2}\ln\frac{|\beta\Sigma_1+(1-\beta)\Sigma_2|}{|\Sigma_1|^{\beta}|\Sigma_2|^{1-\beta}}.$$

Error Bounds

- Chernoff Bound
  - Chernoff bound for P(error) is found by determining the value of β that minimizes $e^{-\kappa(\beta)}$
    - This is a 1-D optimization problem, regardless to dimensionality of class conditional densities.



Error Bounds

- Bhattacharyya Bound
  - The error is given for β=0.5
  - Easier to compute than Chernoff error but looser.

$$P(error) \le \sqrt{P(\omega_1)P(\omega_2)}\int\sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)}\,d\mathbf{x} = \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)},$$

$$k(1/2) = 1/8(\mu_2-\mu_1)^t\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1)+ \frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}.$$

- When the two covariance matrices are equal, k(1/2) is the same as the *Mahalanobis* distance between the two means

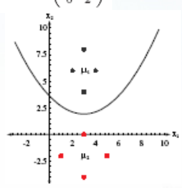
Error Bounds

- Example

$$\mu_1 = \begin{bmatrix}3\\6\end{bmatrix};\quad \Sigma_1 = \begin{pmatrix}1/2 & 0\\0 & 2\end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix}3\\-2\end{bmatrix};\quad \Sigma_2 = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix}2 & 0\\0 & 1/2\end{pmatrix} \quad\text{and}\quad \Sigma_2^{-1} = \begin{pmatrix}1/2 & 0\\0 & 1/2\end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5.$$



- Best Chernoff error bound is 0.008190
- Bhattacharya error bound is 0.008191

ROC

- ROC
- TP, TN, FP, FN
- AUC

Receiver Operating Characteristic (ROC) Curve

- Every classifier employs some kind of a threshold.
  - Changing the threshold affects the performance of the system.
- ROC curves can help us evaluate system performance for different thresholds.
  - Distinguish between discriminability and decision bias, i.e., choice of threshold

Receiver Operating Characteristic (ROC) Curve

- Started in electronic signal detection theory (1940s - 1950s)
  - Has become very popular in biomedical applications, particularly radiology and imaging
  - Also used in machine learning applications to assess classifiers
  - Can be used to compare tests/procedures

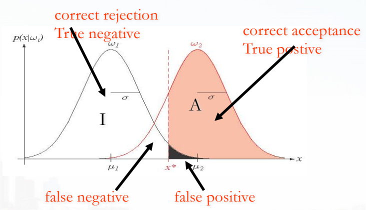
Example: Person Authentication

- Authenticate using biometrics (iPhone5S, …)
  - There are two possible distributions:
    - authentic (A) and impostor (I)

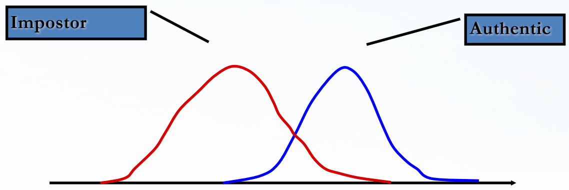
	not rejected	rejected
Authentic (A)	☺ specificity	X Type I error (False +) α
Impostor (I)	X Type II error (False -) β	☺ sensitivity

Example: Person Authentication

- Authenticate a person using biometrics
  - There are two possible distributions:
    - authentic (A) and impostor (I)

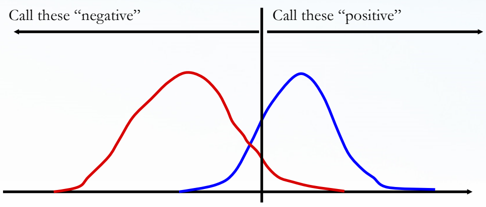


Specific Example



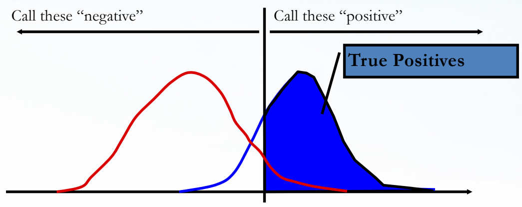
Test Result

Threshold

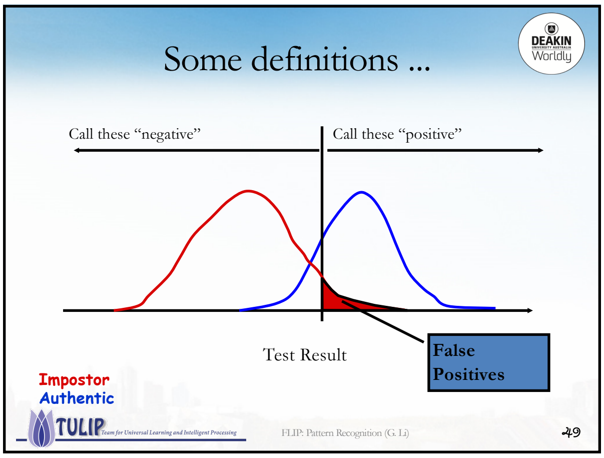


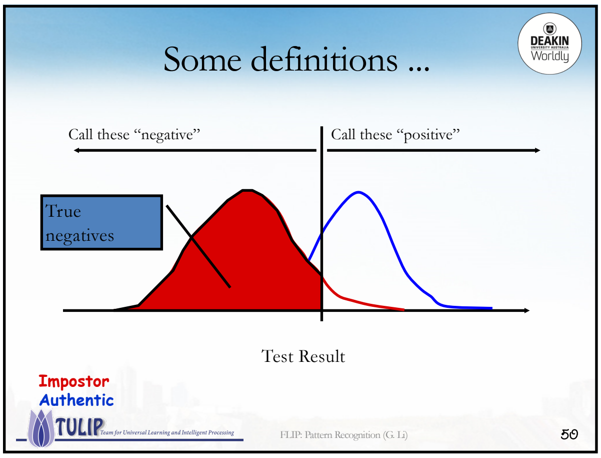
Test Result

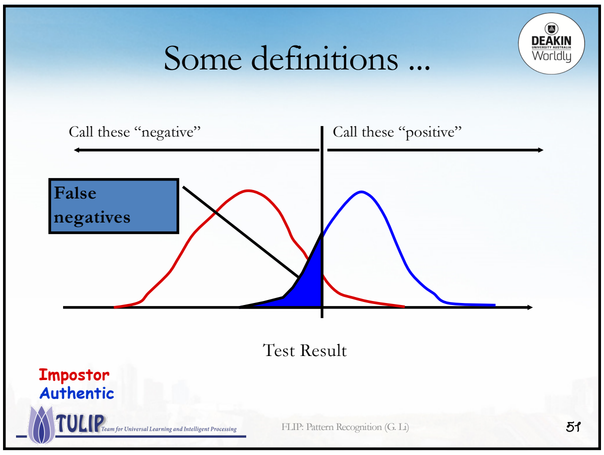
Some definitions ...

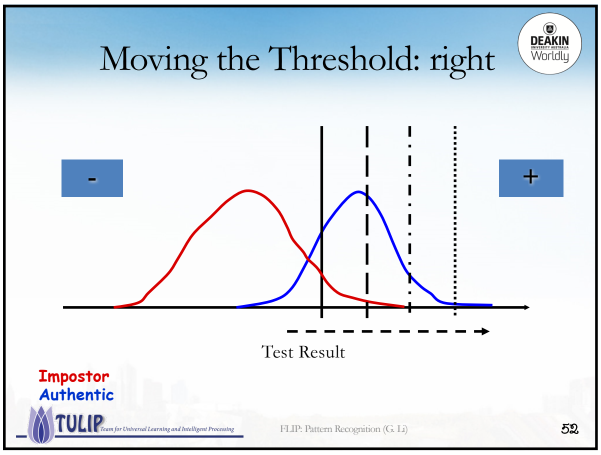


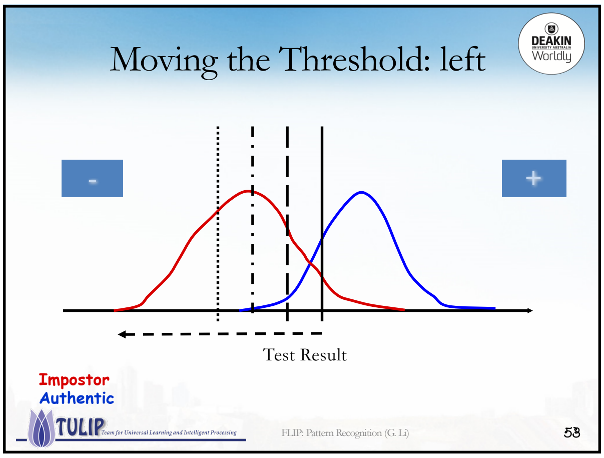
Test Result

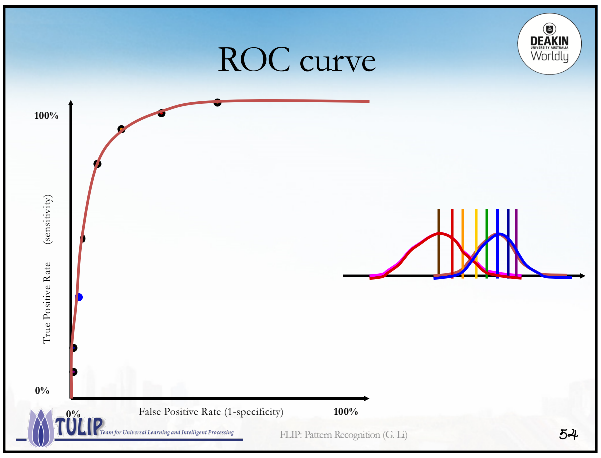

Some definitions ...
Call these "negative"
Call these "positive"
Test Result
False Positives
Impostor
Authentic
TULIP
FLIP: Pattern Recognition (G. Li)
49


Some definitions ...
Call these "negative"
Call these "positive"
True negatives
Test Result
Impostor
Authentic
TULIP
FLIP: Pattern Recognition (G. Li)
50


Some definitions ...
Call these "negative"
Call these "positive"
False negatives
Test Result
Impostor
Authentic
TULIP
FLIP: Pattern Recognition (G. Li)
51


Moving the Threshold: right
-
+
Test Result
Impostor
Authentic
TULIP
FLIP: Pattern Recognition (G. Li)
52

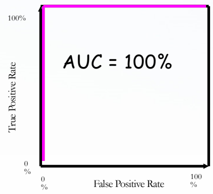

Moving the Threshold: left
-
+
Test Result
Impostor
Authentic
TULIP
FLIP: Pattern Recognition (G. Li)
53

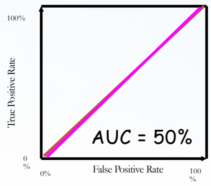

ROC curve
100%
True Positive Rate (sensitivity)
0%
0%
False Positive Rate (1-specificity)
100%
TULIP
FLIP: Pattern Recognition (G. Li)
54

Area under ROC curve (AUC)

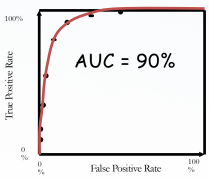
- ***Overall*** measure of test performance
- ***Comparisons*** between two tests based on differences between (estimated) AUC
- For continuous data, AUC equivalent to ***Mann-Whitney U-statistic***
  - nonparametric test of difference in location between two populations

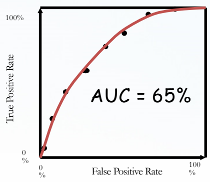
Area under ROC curve (AUC)





Area under ROC curve (AUC)





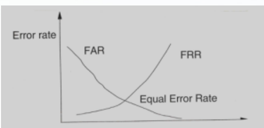
Interpretation of AUC

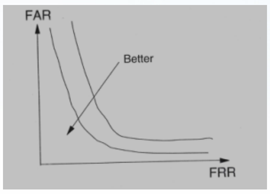
- AUC can be interpreted as the probability that ***the test result from a randomly chosen authentic individual is more indicative of authentic than that from a randomly chosen impostor individual***:

$$P(X_i \geq X_j \mid D_i = 1, D_j = 0)$$

  - So can think of this as a nonparametric distance between authentic/impostor test results

False Positive Rate vs False Negative Rate





Misc

- Discrete Features
- Missing Features
- Noisy Features

Discrete Features

- Components of x are binary or integer valued
  - x can take only one of m discrete values $v_1, v_2, \ldots, v_m$
  - independent binary features for 2-category problem:
    - Let $x = [x_1, x_2, \ldots, x_d]^t$, where each x_i is either 0 or 1, with probabilities:
      - $p_i = P(x_i = 1 \mid \omega_1)$
      - $q_i = P(x_i = 1 \mid \omega_2)$

Discrete Features

- The discriminant function in this case is:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

$where:$

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d$$

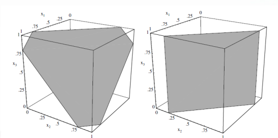
$and:$

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

$decide\ \omega_1$ if $g(x) > 0$ and ω_2 if $g(x) \le 0$

Discrete Features

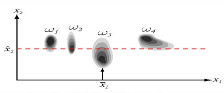
- Consider a 2-class problem with 3 independent binary features
  - class priors are equal and
  - $p_i = 0.8$ and $q_i = 0.5$, $i = 1,2,3;$
  - $w_i = 1.3863$, $w_0 = 1.2$



- Left figure shows the case when p_i=0.8 and q_i=0.5
- Right figure shows case when p_3=q_3

Missing Features

- Suppose it is not possible to measure a certain feature for a given pattern
  - Possible solutions:
    - Reject the pattern
    - Approximate the missing feature
      - **Mean of all the available values for the missing feature**
      - Marginalize over the distribution of the missing feature



- If we set x_1 equal to the average value, we will classify x as ω_3
- But $p(\hat{x}_2 / \omega_2)$ is larger; maybe classify x as ω_2 ?

Missing Features

- Suppose it is not possible to measure a certain feature for a given pattern
  - Possible solutions:
    - Reject the pattern
    - Approximate the missing feature
      - **Mean of all the available values for the missing feature**
      - Marginalize over the distribution of the missing feature

$$P(\omega_i/\mathbf{x}_g) = \frac{P(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int P(\omega_i, \mathbf{x}_g, \mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)} = \frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b)p(\mathbf{x}_g, \mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)} = \frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b)p(\mathbf{x})d\mathbf{x}_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

Decide ω_1 if $P(\omega_1/\mathbf{x}_g) > P(\omega_2/\mathbf{x}_g)$; otherwise decide ω_2

Noisy Features

- Suppose
  - $\mathbf{x}=[\mathbf{x}_g,\mathbf{x}_b]$ ($\mathbf{x}_g$: good features, $\mathbf{x}_b$: noisy features)
  - the noise model $p(\mathbf{x}_b / \mathbf{x}_t)$: $\mathbf{x}_b$: noisy values, $\mathbf{x}_t$: true feature values.
  - statistically independent noise: $\mathbf{x}_b$ is independent of $\mathbf{x}_g$, ω_i

$$P(\omega_i / \mathbf{x}_g, \mathbf{x}_b) = \frac{p(\omega_i, \mathbf{x}_g, \mathbf{x}_b)}{p(\mathbf{x}_g, \mathbf{x}_b)} = \frac{\int P(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)d\mathbf{x}_t}{p(\mathbf{x}_g, \mathbf{x}_b)} = \frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)d\mathbf{x}_t} =$$

$$\frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)p(\mathbf{x}_b/\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_b/\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)d\mathbf{x}_t} = \frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b/\mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_b/\mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)d\mathbf{x}_t} =$$

$$\frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b/\mathbf{x}_t)p(\mathbf{x})d\mathbf{x}_t}{\int p(\mathbf{x}_b/\mathbf{x}_t)p(\mathbf{x})d\mathbf{x}_t}$$

Decide ω_1 if $P(\omega_1/\mathbf{x}_g, \mathbf{x}_b) > P(\omega_2/\mathbf{x}_g, \mathbf{x}_b)$; otherwise decide ω_2

Questions