# Inf2b Learning and Data

## Lecture 10: Discrimination functions

*Hiroshi Shimodaira*
*(Credit: Iain Murray and Steve Renals)*

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2014

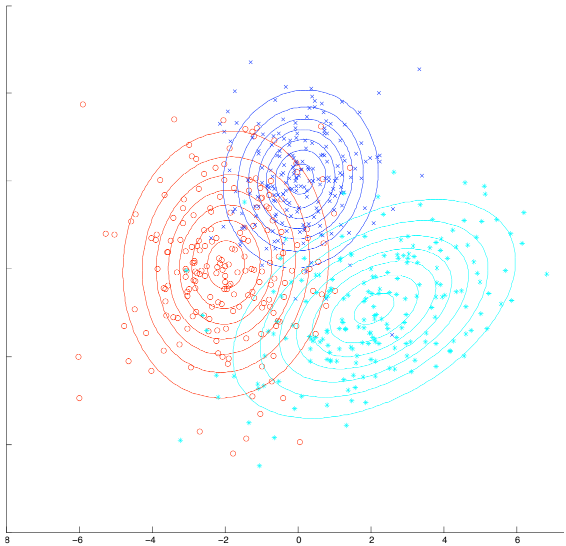
# Today's Schedule

1. Decision Regions
2. Decision Boundaries for minimum error rate classification
3. Deiscriminant Functions

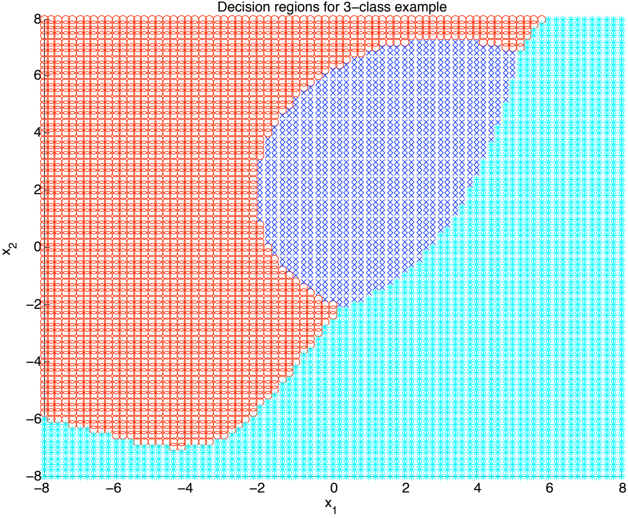
# Decision regions

- Recall Bayes Rule:
$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{p(\mathbf{x})}$$
- Given an unseen point **x**, we assign to the class for which $P(c_i|x)$ is largest.
- Thus **x**-space (the input space) may be regarded as being divided into decision regions $\mathcal{R}_i$ such that a point falling in $\mathcal{R}_i$ is assigned to class $c_i$.
- Decision region $\mathcal{R}_i$ need not be contiguous, but may consist of several disjoint regions each associated with class $c_i$.
- The boundaries between these regions are called decision boundaries

# Gaussians estimated from data



# Decision Regions

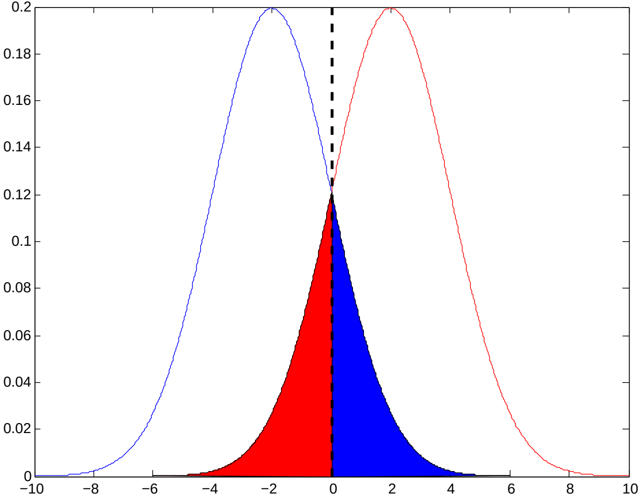


# Placement of decision boundaries

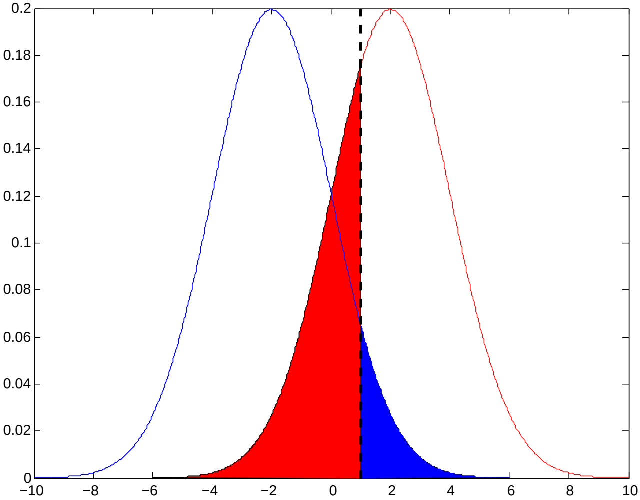
- Consider a 1-dimensional feature space (x) and two classes $c_1$ and $c_2$.
- How to place the decision boundary to minimize the probability of misclassification?
- Misclassification errors $P(\mathrm{error}|x)$:
  1. assigning $x$ to $c_2$ when it belongs to $c_1$ ($x$ is in $\mathcal{R}_2$ when it belongs to $c_1$) $\cdots P(c_1|x)$
  2. assigning $x$ to $c_1$ when it belongs to $c_2$ ($x$ is in $\mathcal{R}_1$ when it belongs to $c_2$) $\cdots P(c_2|x)$
- Total probability of error:

$$
\begin{aligned}
P(\mathrm{error}) &= \int P(\mathrm{error}|x) p(x) dx = P(x \in \mathcal{R}_2, c_1) + P(x \in \mathcal{R}_1, c_2) \\
&= P(x \in \mathcal{R}_2 | c_1) P(c_1) + P(x \in \mathcal{R}_1 | c_2) P(c_2) \\
&= \int_{\mathcal{R}_2} p(x\,|\,c_1)\, P(c_1)\, \mathrm{d}x + \int_{\mathcal{R}_1} p(x\,|\,c_2)\, P(c_2)\, \mathrm{d}x
\end{aligned}
$$

# Decision boundaries and misclassification



# Decision boundaries and misclassification



# Minimising probability of misclassification

$$P(\text{error}) = \int_{\mathcal{R}_2} p(x \,|\, c_1)\, P(c_1)\, \mathrm{d}x + \int_{\mathcal{R}_1} p(x \,|\, c_2)\, P(c_2)\, \mathrm{d}x$$

- To minimise $P(\text{error})$:
  For a given $x$ if $p(x|c_1)P(c_1) > p(x|c_2)P(c_2)$, then point $x$ should be in region $\mathcal{R}_1$
- The probability of misclassification is thus minimised by assigning each point to the class with the maximum posterior probability (Bayes decision rule / MAP decision rule / minimum error rate classification)
- This justification for the maximum posterior probability may be extended to d-dimensional feature vectors and $K$ classes

1. Decision Regions

2. Decision Boundaries for minimum error rate classification

3. Deiscriminant Functions

# Discriminant functions

- We can express a classification rule in terms of a discriminant function $y_c(\mathbf{x})$ for each class, such that x is assigned to class $c$ if:

$$y_c(\mathbf{x}) > y_k(\mathbf{x}) \quad \forall\ k \neq c$$

- If we assign $\mathbf{x}$ to class $c$ with the highest posterior probability $P(c|\mathbf{x})$, then the posterior probability or the log posterior probability forms a suitable discriminant function:

$$y_c(\mathbf{x}) = \ln P(C\,|\,\mathbf{x}) \ \propto\ \ln p(\mathbf{x}\,|\,c) + \ln P(c)$$

- Decision boundaries are defined when the discriminant functions are equal: $y_k(\mathbf{x}) = y_\ell(\mathbf{x})$
- Decision boundaries are not changed by monotonic transformations (such as taking the log) of the discriminant functions.

# Discriminant functsions for Gaussian pdfs

- What is the form of the discriminant function when using a Gaussian pdf?
- If the discriminant function is the log posterior probability:

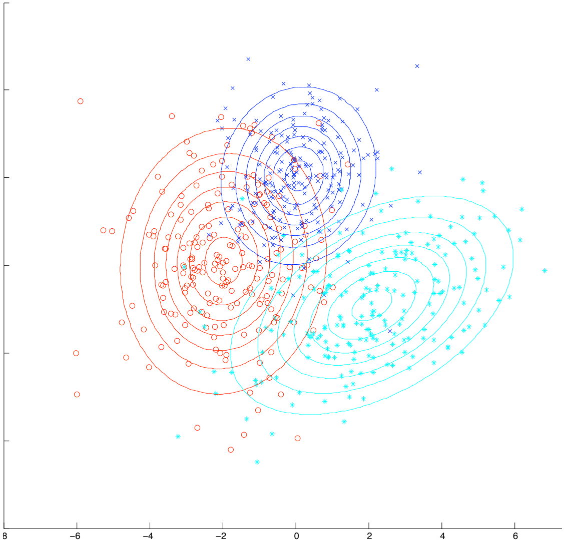
$$y_c(\mathbf{x}) = \ln p(\mathbf{x}|C) + \ln P(C)$$

- Then, substituting in the log probability of a Gaussian and dropping constant terms we obtain:

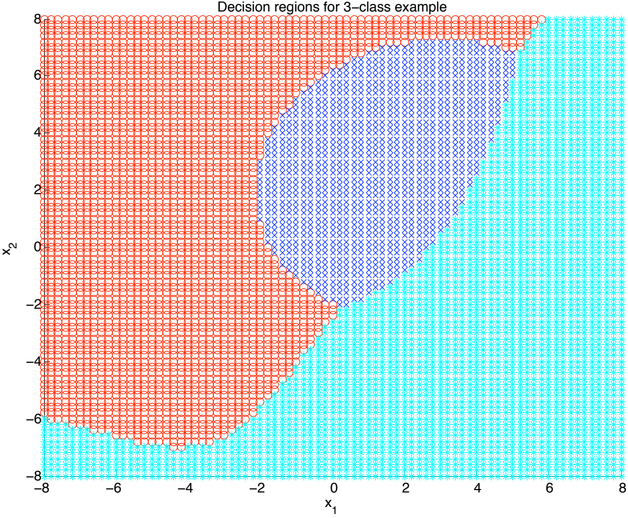
$$y_c(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_c| + \ln P(C)$$

- This function is quadratic in $\mathbf{x}$

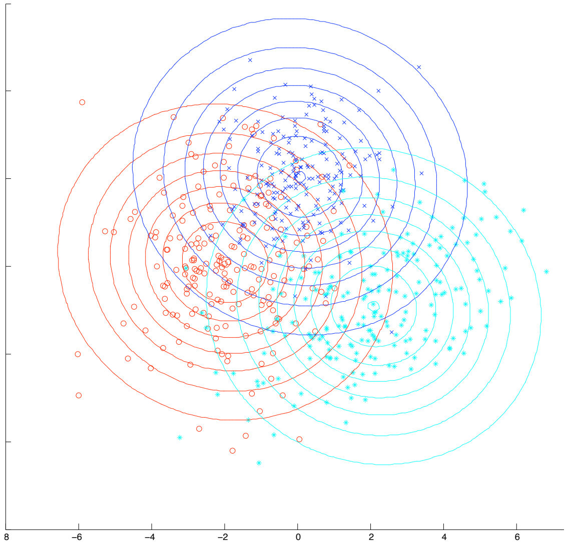
# Gaussians estimated from training data



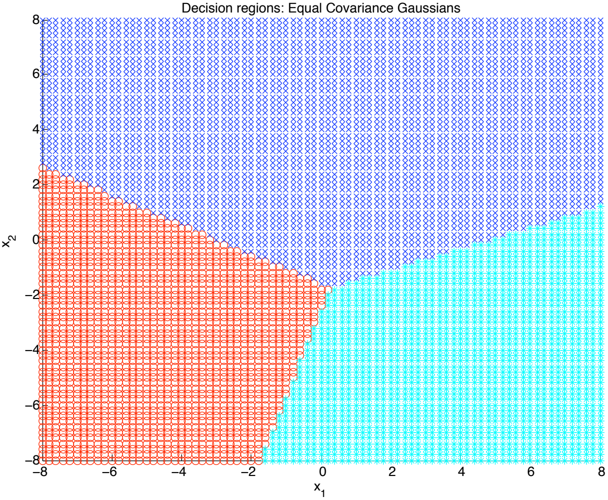
# Decision Regions



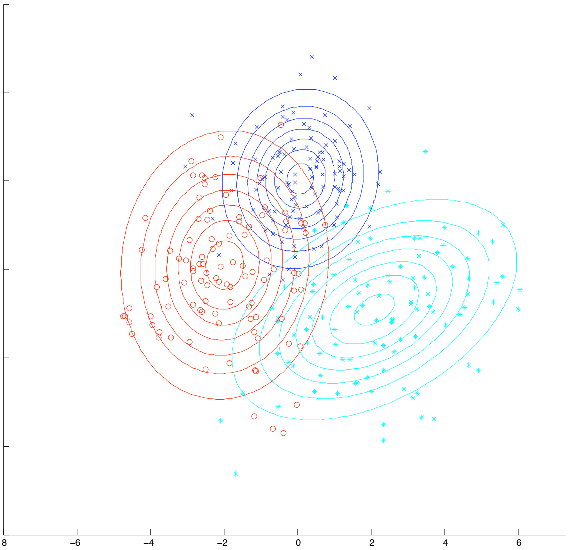
# Equal Covariance Gaussians estimated from the data



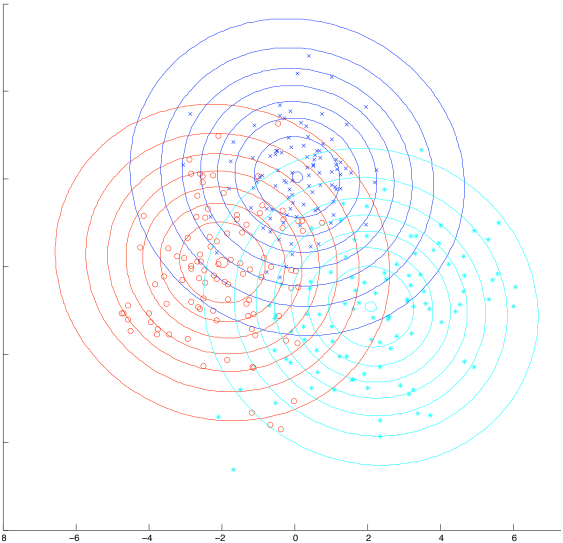
# Decision Regions: Σ shared



# Testing data (Non-equal covariance)



# Testing data (Equal covariance)



# Results

- Non-equal covariance Gaussians

| Test Data | | True class A | B | C |
|---|---|---|---|---|
| Predicted | A | 77 | 5 | 9 |
| class | B | 15 | 88 | 2 |
| | C | 8 | 7 | 89 |

Fraction correct: $(77 + 88 + 89)/300 = 254/300 = 0.85$.

- Equal covariance Gaussians

| Test Data | | True class A | B | C |
|---|---|---|---|---|
| Predicted | A | 80 | 10 | 8 |
| class | B | 14 | 90 | 6 |
| | C | 6 | 0 | 86 |

Fraction correct: $(80 + 90 + 86)/300 = 256/300 = 0.85$.

# Gaussians with equal covariance

- Consider the special case in which the Gaussian pdfs for each class all share the same class-independent covariance matrix: $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma},\ \forall\, c$

$$y_c(\mathbf{x})^{(org)} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(c)$$

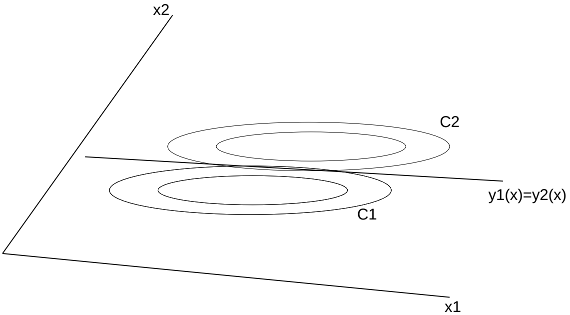
$$\begin{aligned} y_c(\mathbf{x}) &= \left(\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}^{-1}\right)\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \ln P(c) \\ &= \mathbf{w}_c^T \mathbf{x} + w_{c0} \end{aligned}$$

where

$$\mathbf{w}_c^T = \boldsymbol{\mu}_c^T \boldsymbol{\Sigma}^{-1}, \quad w_{c0} = -\frac{1}{2}\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \ln P(c)$$

- This is called a linear discriminant function, as it is a linear function of $\mathbf{x}$.

# Linear discriminant: decision boundary for equal covariance Gaussians




- In two dimensions the boundary is a line
- In three dimensions it is a plane
- In $d$ dimensions it is a hyperplane
  (i.e. $\{\mathbf{x} \mid \mathbf{w}_c^T\mathbf{x} + w_{c0} = 0\}$)

# Spherical Gaussians with Equal Covariance

- Spherical Gaussians have a diagonal covariance matrix, with the same variance in each dimension

$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2} \mathbf{I}$$

- If we further assume that the prior probabilities of each class are equal, we can write the discriminant function as

$$y_c(\mathbf{x}) = -\frac{||\mathbf{x} - \boldsymbol{\mu}_c||^2}{2\sigma^2} + \ln P(c)$$

- If the prior probabilities are eqaul for all classes, the decision rule: "assign a test data to the class whose mean is closest".

  In this case the class means ($\boldsymbol{\mu}_c$) may be regarded as class templates or prototypes.

# Two-clas linear discriminants

- For a two class problem, the log odds can be used as a single discriminant function:

$$\begin{aligned} y(\mathbf{x}) &= \ln \frac{P(c_1 \,|\, \mathbf{x})}{P(c_2 \,|\, \mathbf{x})} = \ln \frac{p(\mathbf{x} \,|\, c_1)\, P(c_1)}{p(\mathbf{x} \,|\, c_2)\, P(c_2)} \\ &= \ln p(\mathbf{x} \,|\, c_1) - \ln p(\mathbf{x} \,|\, c_2) + \ln P(c_1) - \ln P(c_2) \end{aligned}$$

- If the pdf is a Gaussian with the shared covariance matrix, we have a linear discriminant:

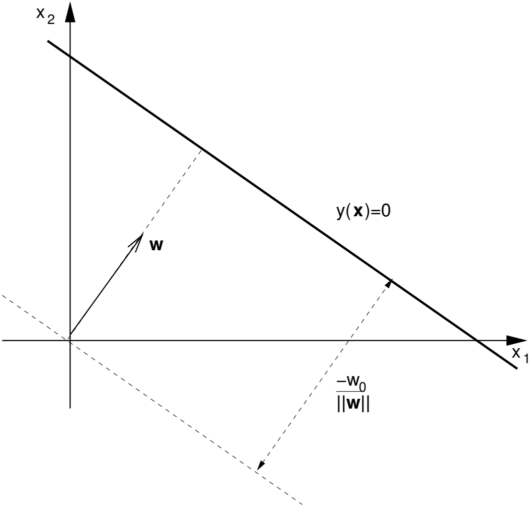
  $$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

  $\mathbf{w}$ and $w_0$ are functions of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma, P(c_1)$, and $P(c_2)$.
- Let $\mathbf{x}_a$ and $\mathbf{x}_b$ be two points on the decision boundary

  $$\mathbf{w}^T \mathbf{x}_a + w_0 = \mathbf{w}^T \mathbf{x}_b + w_0 = 0$$

  $$\mathbf{w}^T (\mathbf{x}_a - \mathbf{x}_b) = 0, \quad i.e.\ \mathbf{w} \perp (\mathbf{x}_a - \mathbf{x}_b)$$

# Geometry of a two-class linear discriminant




- $\mathbf{w}$ is normal to any vector on the hyperplane decision boundary
- If $\mathbf{x}$ is a point on the hyperplane, then the normal

# Summary

- Obtaining decision boundaries from probability models and a decision rule
- Minimising the probability of error
- Discriminant functions and Gaussian pdfs
- Linear discriminants and Gaussians with equal covariance
- There are many other ways to train discriminants