Inf2b - Learning

Lecture 10: Discriminant functions

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

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

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://piazza.com/ed.ac.uk/spring2020/infr08028
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's Schedule

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

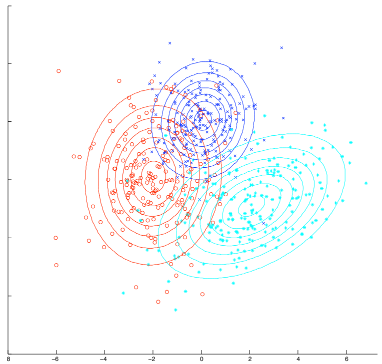
Decision regions

- Recall Bayes' Rule:

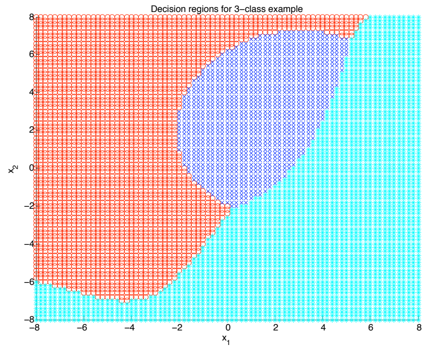
$$P(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x})}$$

- Given an unseen point $\boldsymbol{x}$, we assign to the class for which $P(C_k|x)$ is largest. ($k^* = \arg\max_k P(C_k|x)$)
- Thus $\boldsymbol{x}$-space (the input space) may be regarded as being divided into decision regions $\mathcal{R}_k$ such that a point falling in $\mathcal{R}_k$ is assigned to class C_k.
- Decision region $\mathcal{R}_k$ need not be contiguous, but may consist of several disjoint regions each associated with class C_k.
- The boundaries between these regions are called decision boundaries. (Recall the examples of decision boundaries by k-NN classification in Chapter 4)

Gaussians estimated from data

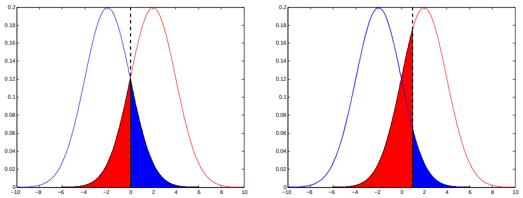


Decision Regions



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 minimise the probability of misclassification (based on $p(x, C_k)$)?



Decision regions and misclassification

Confusion matrix

In\Out	C_1	C_2
C_1	N_{11}	N_{12}
C_2	N_{21}	N_{22}

$\Rightarrow$

Normalised version

In\Out	C_1	C_2	
C_1	P_{11}	P_{12}	$P_{11}+P_{12}=1$
C_2	P_{21}	P_{22}	$P_{21}+P_{22}=1$

$$P_{11} = P(x \in \mathcal{R}_1|C_1) = \tfrac{N_{11}}{N_1}, \quad P_{12} = P(x \in \mathcal{R}_2|C_1) = \tfrac{N_{12}}{N_1}$$
$$P_{21} = P(x \in \mathcal{R}_1|C_2) = \tfrac{N_{21}}{N_2}, \quad P_{22} = P(x \in \mathcal{R}_2|C_2) = \tfrac{N_{22}}{N_2}$$

$N_1 = N_{11}+N_{12},\ N_2 = N_{21}+N_{22},\ P(C_1) = \frac{N_1}{N_1+N_2},\ P(C_2) = \frac{N_2}{N_1+N_2}$

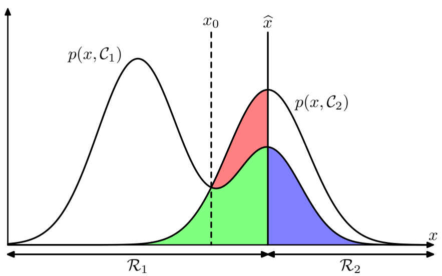
$$P(\text{correct}) = \frac{N_{11}+N_{22}}{N_1+N_2} = P_{11}\,P(C_1) + P_{22}\,P(C_2)$$
$$P(\text{error}) = \frac{N_{12}+N_{21}}{N_1+N_2} = P_{12}\,P(C_1) + P_{21}\,P(C_2)$$
$$= \int_{\mathcal{R}_2} p(x|C_1)\,P(C_1)\,\mathrm{d}x + \int_{\mathcal{R}_1} p(x|C_2)\,P(C_2)\,\mathrm{d}x$$

Minimising probability of misclassification

$$P(\text{error}|\mathcal{R}_1, \mathcal{R}_2) = \int_{\mathcal{R}_2} p(x\,|\,C_1)\,P(C_1)\,\mathrm{d}x + \int_{\mathcal{R}_1} p(x\,|\,C_2)\,P(C_2)\,\mathrm{d}x$$

- If there is $x_e \in \mathcal{R}_2$ such that $p(x_e|C_1)P(C_1) > p(x_e|C_2)P(C_2)$, letting $\mathcal{R}_2^* = \mathcal{R}_2 - \{x_e\}$ and $\mathcal{R}_1^* = \mathcal{R}_1 + \{x_e\}$ gives
$$P(\text{error}|\mathcal{R}_1^*, \mathcal{R}_2^*) < P(\text{error}|\mathcal{R}_1, \mathcal{R}_2)$$
- $P(\text{error})$ is 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

Minimising probability of misclassification *(cont.)*



After Fig. 1.24, C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
$\hat{x}$ denotes the current decision boundary, which causes error shown in red, green, and blue regions. The error is minimised by locating the boundary at x_o.

Discriminant functions

- We can express a classification rule in terms of a discriminant function $y_k(x)$ for each class, such that x is assigned to class C_k if:

$$y_k(x) > y_\ell(x) \quad \forall \ell \neq k$$

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

$$y_k(\mathbf{x}) = \ln p(\mathbf{x}|\,C_k) + \ln P(C_k)$$

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

Discriminant functions for Gaussian pdfs

- What is the form of the discriminant function when using a Gaussian pdf?

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

- If the discriminant function is the log posterior probability:

$$y_k(x) = \ln p(\mathbf{x}|C_k) + \ln P(C_k)$$

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

$$y_k(x) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln P(C_k)$$

- This function is quadratic in x

Discriminant functions for Gaussian pdfs *(cont.)*

- To see if the function is really quadratic in x,

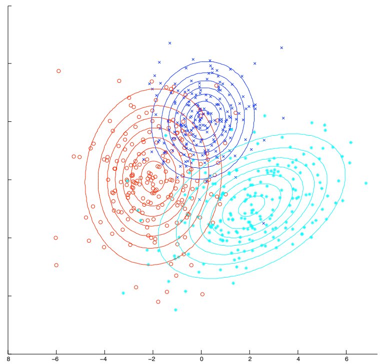
$$\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \\
&= \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} - \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} - \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\
&= \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} - 2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k
\end{aligned}$$

- In 2-D case, let $\boldsymbol{\Sigma}_k^{-1} = A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$,

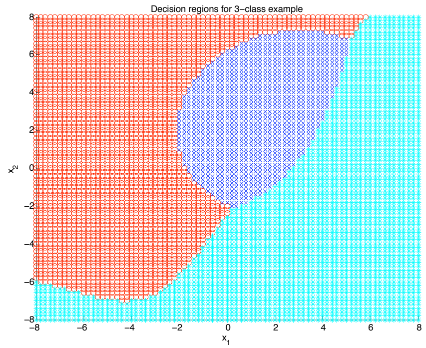
$$\begin{aligned}
\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} &= \mathbf{x}^T A\,\mathbf{x} \\
&= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= a_{11}x_1^2 + (a_{12}+a_{21})x_1x_2 + a_{22}x_2^2
\end{aligned}$$

See Note 10 for details.

Gaussians estimated from training data



Decision Regions



Gaussians with equal covariance

$$\begin{aligned}
y_k(\mathbf{x}) &= -\frac{1}{2}(x-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(x-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln P(C_k) \\
&= -\frac{1}{2}(x^T\boldsymbol{\Sigma}_k^{-1}x - 2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}x + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln P(C_k)
\end{aligned}$$

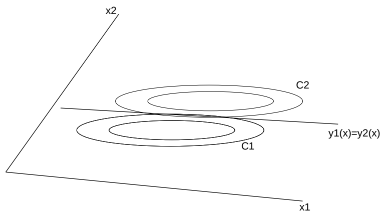
- Consider the special case in which the Gaussian pdfs for each class all share the same class-independent covariance matrix: $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}, \ \forall\, C_k$

$$\begin{aligned}
y_k(\mathbf{x}) &= \left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\right)\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln P(C_k) \\
&= \mathbf{w}_k^T\mathbf{x} + w_{k0} \qquad = w_{kD}x_D + \cdots + w_{k1}x_1 + w_{k0}
\end{aligned}$$

where $\quad \mathbf{w}_k^T = \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}, \quad w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln P(C_k)$

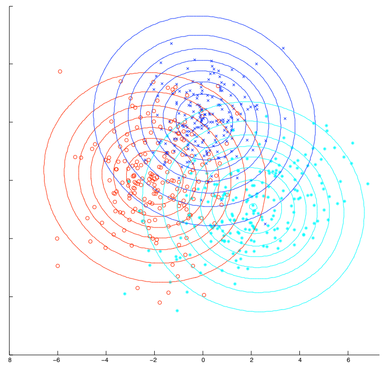
- This is called a linear discriminant function, as it is a linear function of $\mathbf{x}$.

Gaussians with equal covariance *(cont.)*

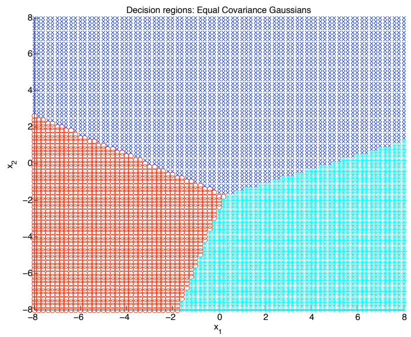


- In two dimensions the boundary is a line
- In three dimensions it is a plane
- In D dimensions it is a hyperplane (i.e. $\{\mathbf{x} \mid \mathbf{w}_k^T\mathbf{x} + w_{k0} = 0\}$)

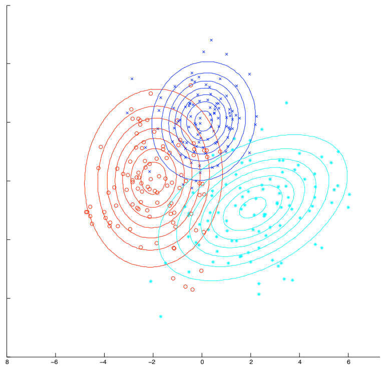
Gaussians estimated from the data: Σ shared



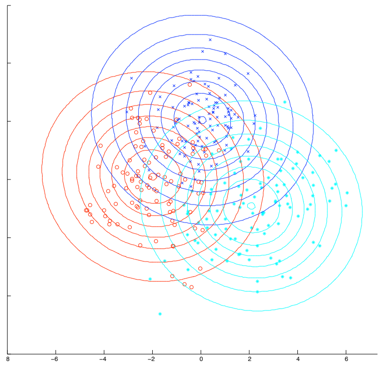
Decision Regions: Σ shared



Testing data (Non-equal covariance)



Testing data (Equal covariance)



Results

- Non-equal covariance Gaussians

Test Data		*Predicted class* A	B	C
Actual	A	77	15	8
class	B	5	88	7
	C	9	2	89

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

- Equal covariance Gaussians

Test Data		*Predicted class* A	B	C
Actual	A	80	14	6
class	B	10	90	0
	C	8	6	86

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

Spherical Gaussians with Equal Covariance

- Spherical Gaussians: $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$

$$\Rightarrow \quad |\boldsymbol{\Sigma}| = \sigma^{2D}, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbf{I}$$

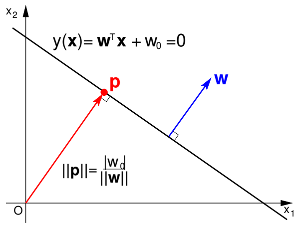
$$y_k(\mathbf{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln P(C_k)$$
$$= -\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln\sigma^{2D} + \ln P(C_k)$$
$$y_k(\mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2 + \ln P(C_k)$$

- If equal prior probabilities are assumed,
$$y_k(\mathbf{x}) = -\|\mathbf{x}-\boldsymbol{\mu}_k\|^2$$
The decision rule: "assign a test data to the class whose mean is closest".
The class means ($\boldsymbol{\mu}_k$) may be regarded as class templates or prototypes.

Two-class linear discriminants

- For a two class problem, the log odds can be used as a single discriminant function:
$$y(\mathbf{x}) = \ln\frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} = \ln\frac{p(\mathbf{x}|C_1)P(C_1)}{p(\mathbf{x}|C_2)P(C_2)}$$
$$= \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln P(C_1) - \ln P(C_2)$$
- 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 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}, P(C_1)$, and $P(C_2)$.
- $\mathbf{w}$ is a normal vector to the decision boundary.
Let $\mathbf{a}$ and $\mathbf{b}$ be two points on the decision boundary
$$\mathbf{w}^T\mathbf{a} + w_0 = \mathbf{w}^T\mathbf{b} + w_0 = 0 \quad \Rightarrow \quad \mathbf{w}^T(\mathbf{a}-\mathbf{b}) = 0$$
$$i.e.\ \ \mathbf{w} \perp (\mathbf{a}-\mathbf{b})$$

Geometry of a two-class linear discriminant



- $\mathbf{w}$ is normal to the decision boundary (hyperplane), $\mathbf{w}^T\mathbf{x} + w_0 = 0$.
- If $\mathbf{p}$ is the point on the hyperplane closest to the origin, then the normal distance from the hyperplane to the origin is given by:
$$\|p\| = \frac{\mathbf{w}^T\mathbf{p}}{\|\mathbf{w}\|} = \frac{|w_0|}{\|\mathbf{w}\|}$$

$$0 = \mathbf{w}^T\mathbf{p} + w_0$$
$$= \|\mathbf{w}\|\|\mathbf{p}\|\cos 0 + w_0$$
$$= \|\mathbf{w}\|\|\mathbf{p}\| \pm w_0$$

Exercise

1. Considering a classification problem of two classes, where each class is modelled with a D-dimensional Gaussian distribution. Derive the formula for the decision boundary, and show that it is quadratic in $\mathbf{x}$.
2. Considering a classification problem of two classes, whose discriminant function takes the form, $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$.
   - Confirm that the decision boundary is a straight line when $D = 2$.
   - Confirm that the weight vector $\mathbf{w}$ is a normal vector to the decision boundary.
3. Try Lab-7 on Classification with Gaussians

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
- In next lectures, we will see discriminant functions trained with different criteria.