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
Recall Bayes’ Rule:

\[ P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)} \]

Given an unseen point \( x \), we assign to the class for which \( P(C_k|x) \) is largest. (\( k^* = \text{arg max}_k P(C_k|x) \))

Thus \( x \)-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

Inf2b - Learning: Lecture 10  Discriminant functions
Decision Regions

Decision regions for 3-class example

Inf2b - Learning: Lecture 10  Discriminant functions  5
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

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>Normalised version</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{In} )</td>
<td>( \text{Out} )</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>( N_{11} )</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>( N_{21} )</td>
</tr>
</tbody>
</table>

\[ P_{11} = P(x \in \mathcal{R}_1 | C_1) = \frac{N_{11}}{N_1}, \quad P_{12} = P(x \in \mathcal{R}_2 | C_1) = \frac{N_{12}}{N_1} \]

\[ P_{21} = P(x \in \mathcal{R}_1 | C_2) = \frac{N_{21}}{N_2}, \quad P_{22} = P(x \in \mathcal{R}_2 | C_2) = \frac{N_{22}}{N_2} \]

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

\[ P(\text{correct}) = \frac{N_{11} + N_{22}}{N_1 + N_2} = P_{11} P(C_1) + P_{22} P(C_2) \]

\[ P(\text{error}) = \frac{N_{12} + N_{21}}{N_1 + N_2} = P_{12} P(C_1) + P_{21} P(C_2) \]

\[ = \int_{\mathcal{R}_2} p(x|C_1) P(C_1) \, dx + \int_{\mathcal{R}_1} p(x|C_2) P(C_2) \, dx \]
Minimising probability of misclassification

\[ P(\text{error}|\mathcal{R}_1, \mathcal{R}_2) = \int_{\mathcal{R}_2} p(x|C_1) P(C_1) \, dx + \int_{\mathcal{R}_1} p(x|C_2) P(C_2) \, dx \]

- 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.)

\[ \hat{x} \]

\[ p(x, C_1) \]
\[ p(x, C_2) \]

\[ x \]
\[ R_1 \]
\[ R_2 \]

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 \( x_0 \).
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 $P(C|x)$, then the log posterior probability forms a suitable discriminant function:

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

Decision boundaries between $C_k$ and $C_\ell$ 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.
What is the form of the discriminant function when using a Gaussian pdf?

\[ p(x | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) \right) \]

If the discriminant function is the log posterior probability:

\[ y_k(x) = \ln p(x | C_k) + \ln P(C_k) \]

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

\[ y_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \ln |\Sigma_k| + \ln P(C_k) \]

This function is quadratic in \( x \)
To see if the function is really quadratic in \( x \),

\[
(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)
\]

\[
= x^T \Sigma_k^{-1} x - \mu_k^T \Sigma_k^{-1} x - x^T \Sigma_k^{-1} \mu_k + \mu_k^T \Sigma_k^{-1} \mu_k
\]

\[
= x^T \Sigma_k^{-1} x - 2 \mu_k^T \Sigma_k^{-1} x + \mu_k^T \Sigma_k^{-1} \mu_k
\]

In 2-D case, let \( \Sigma_k^{-1} = A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \),

\[
x^T \Sigma_k^{-1} x = x^T A x
\]

\[
= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}
\]

\[
= a_{11} x_1^2 + (a_{12} + a_{21}) x_1 x_2 + a_{22} x_2^2
\]

See Note 10 for details.
Gaussians estimated from training data
Decision Regions

Decision regions for 3-class example

Discriminant functions
Gaussians with equal covariance

\[ y_k(x) = -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) - \frac{1}{2} \ln |\Sigma_k| + \ln P(C_k) \]

\[ = -\frac{1}{2} (x^T \Sigma_k^{-1} x - 2 \mu_k^T \Sigma_k^{-1} x + \mu_k^T \Sigma_k^{-1} \mu_k) - \frac{1}{2} \ln |\Sigma_k| + \ln P(C_k) \]

- Consider the special case in which the Gaussian pdfs for each class all share the same class-independent covariance matrix: \( \Sigma_k = \Sigma, \ \forall C_k \)

\[ y_k(x) = (\mu_k^T \Sigma^{-1}) x - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln P(C_k) \]

\[ = w_k^T x + w_{k0} \]

where \( w_k^T = \mu_k^T \Sigma^{-1}, \quad w_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln P(C_k) \)

- This is called a linear discriminant function, as it is a linear function of \( x \).
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: $\Sigma$ shared
Decision Regions: $\Sigma$ shared

Decision regions: Equal Covariance Gaussians

Inf2b - Learning: Lecture 10 Discriminant functions
Testing data (Non-equal covariance)
Testing data (Equal covariance)
Results

- Non-equal covariance Gaussians

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>77</td>
</tr>
<tr>
<td>class B</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
</tr>
</tbody>
</table>

Fraction correct: \( \frac{77 + 88 + 89}{300} = \frac{254}{300} \approx 0.85 \).

- Equal covariance Gaussians

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>80</td>
</tr>
<tr>
<td>class B</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
</tr>
</tbody>
</table>

Fraction correct: \( \frac{80 + 90 + 86}{300} = \frac{256}{300} \approx 0.85 \).
Spherical Gaussians with Equal Covariance

- Spherical Gaussians: \( \Sigma = \sigma^2 I \)

\[
\Rightarrow |\Sigma| = \sigma^{2D}, \quad \Sigma^{-1} = \frac{1}{\sigma^2} I
\]

\[
y_k(x) = -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) - \frac{1}{2} \ln |\Sigma_k| + \ln P(C_k)
\]

\[
= -\frac{1}{2\sigma^2} (x - \mu_k)^T (x - \mu_k) - \frac{1}{2} \ln \sigma^{2D} + \ln P(C_k)
\]

\[
y_k(x) = -\frac{1}{2\sigma^2} \|x - \mu_k\|^2 + \ln P(C_k)
\]

- If equal prior probabilities are assumed,

\[
y_k(x) = -\|x - \mu_k\|^2
\]

The decision rule: “assign a test data to the class whose mean is closest”.

The class means (\(\mu_k\)) 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:

\[ y(x) = \ln \frac{P(C_1 | x)}{P(C_2 | x)} = \ln \frac{p(x | C_1) P(C_1)}{p(x | C_2) P(C_2)} \]

\[ = \ln p(x | C_1) - \ln p(x | C_2) + \ln P(C_1) - \ln P(C_2) \]

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

\[ y(x) = w^T x + w_0 \]

\( w \) and \( w_0 \) are functions of \( \mu_1, \mu_2, \Sigma, P(C_1), \) and \( P(C_2). \)

- \( w \) is a normal vector to the decision boundary.

Let \( a \) and \( b \) be two points on the decision boundary

\[ w^T a + w_0 = w^T b + w_0 = 0 \quad \Rightarrow \quad w^T (a - b) = 0 \]

i.e. \( w \perp (a - b) \)
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:

\[
\| \mathbf{p} \| = \frac{\mathbf{w}^T \mathbf{p}}{\| \mathbf{w} \|} = \frac{|w_0|}{\| \mathbf{w} \|}
\]

\[
0 = \mathbf{w}^T \mathbf{p} + w_0 = \| \mathbf{w} \| \| \mathbf{p} \| \cos \theta + w_0 = \| \mathbf{w} \| \| \mathbf{p} \| \pm w_0
\]
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 $x$.

2. Considering a classification problem of two classes, whose discriminant function takes the form, $y(x) = w^T x + w_0$.
   - Confirm that the decision boundary is a straight line when $D = 2$.
   - Confirm that the weight vector $w$ is a normal vector to the decision boundary.

3. Try Lab-7 on Classification with Gaussians
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.