Inf2b - Learning

Lecture 9: Classification with Gaussians

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

Classification with Gaussians

1. The multidimensional Gaussian distribution (recap.)
2. Practical topics on covariance matrix
3. Bayes theorem and probability densities
4. 1-dimensional Gaussian classifier
5. Multivariate Gaussian classifier
6. Evaluation of classifier performance

The multidimensional Gaussian distribution

- The D-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_D)^T$ is multivariate Gaussian if it has a probability density function of the following form:
$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$
The pdf is parameterised by the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ.
- The 1-dimensional Gaussian is a special case of this pdf
- The argument to the exponential $\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is referred to as a *quadratic form*, and it is always *non-negative*.

Covariance matrix

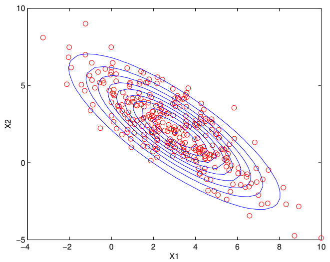
Covariance matrix (with ML estimation):

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1D} \\ \vdots & \ddots & \vdots \\ \sigma_{D1} & \cdots & \sigma_{DD} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T$$

where $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD})^T$
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^T$

- Symmetric : $\Sigma^T = \Sigma$, and $(\Sigma^{-1})^T = \Sigma^{-1}$
- Semi-positive definite: $\boldsymbol{x}^T\Sigma\boldsymbol{x} \ge 0$, and $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} \ge 0$
- cf: sample covariance matrix, which uses $\frac{1}{N-1}$.

Maximum likelihood fit to a Gaussian



Tips on calculating covariance matrices

MATLAB is optimised for matrix/vector operations

$$\underset{(D\times D)}{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\underset{(D\times 1)}{(\boldsymbol{x}_n-\boldsymbol{\mu})}\underset{(1\times D)}{(\boldsymbol{x}_n-\boldsymbol{\mu})^T}$$

$$= \frac{1}{N}\underset{(D\times N)}{(\boldsymbol{x}_1-\boldsymbol{\mu}, \ldots, \boldsymbol{x}_N-\boldsymbol{\mu})}\underset{(N\times D)}{\begin{pmatrix} \boldsymbol{x}_1^T-\boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{x}_N^T-\boldsymbol{\mu}^T \end{pmatrix}}$$

$$= \frac{1}{N}\underset{(D\times N)}{(X-M_N)^T}\underset{(N\times D)}{(X-M_N)}$$

$$\underset{(N\times D)}{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11}, \ldots, x_{1D} \\ \vdots \qquad \vdots \\ x_{N1}, \ldots, x_{ND} \end{bmatrix}, \quad \underset{(N\times D)}{M_N} = \begin{bmatrix} M \\ \vdots \\ M \end{bmatrix} = \begin{bmatrix} \mu_1, \ldots, \mu_D \\ \vdots \qquad \vdots \\ \mu_1, \ldots, \mu_D \end{bmatrix}$$

$$\underset{(1\times D)}{M} = \boldsymbol{\mu}^T = [\ \mu_1, \ldots, \mu_D\ ], \qquad = \tfrac{1}{N}1_{NN}X$$

Properties of covariance matrix

$$\Sigma = V D V^T$$

$$= \begin{pmatrix} v_{11} & \cdots & v_{1D} \\ \vdots & \ddots & \vdots \\ v_{D1} & \cdots & v_{DD} \end{pmatrix}\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_D \end{pmatrix}\begin{pmatrix} v_{11} & \cdots & v_{1D} \\ \vdots & \ddots & \vdots \\ v_{D1} & \cdots & v_{DD} \end{pmatrix}^T$$

$$= (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_D)\ \mathrm{Diag}(\lambda_1, \ldots, \lambda_D)\ (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_D)^T$$

- $\boldsymbol{v}_i$: eigen vector, λ_i : eigen value
$\Sigma\,\boldsymbol{v}_i = \lambda_i\,\boldsymbol{v}_i$
- $\lambda_i \ge 0$, $\|\boldsymbol{v}_i\| = 1$
- $|\Sigma| = \prod_{i=1}^{D}\lambda_i$
- $\sum_{i=1}^{D}\sigma_{ii} = \sum_{i=1}^{D}\lambda_i$

Properties of covariance matrix

- $\mathrm{rank}(\Sigma)$
  - the number of linearly independent columns (or rows)
  - the number of bases (i.e. the dimension of the column space)

$\mathrm{rank}(\Sigma) = D \;\rightarrow\; \forall_i\ :\ \lambda_i > 0$

$\forall_{i\neq j}\ :\ \boldsymbol{v}_i \perp \boldsymbol{v}_j$

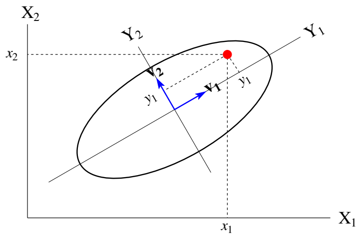
$|\Sigma| > 0$

$\mathrm{rank}(\Sigma) < D \;\rightarrow\; \exists_i\ :\ \lambda_i = 0$

$\exists_{(i,j)}\ :\ \rho(x_i, x_j) = 1$

$|\Sigma| = 0$

Geometry of covariance matrix

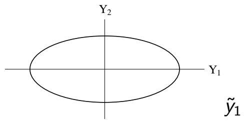


Sort eigen values: $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_D$

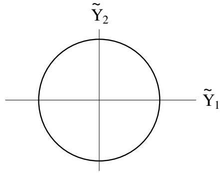
$\boldsymbol{v}_1$: eigen vector of λ_1
$\boldsymbol{v}_2$: eigen vector of λ_2

$y_1 = \boldsymbol{v}_1^T\boldsymbol{x}$, $\mathrm{Var}(y_1) = \lambda_1$
$y_2 = \boldsymbol{v}_2^T\boldsymbol{x}$, $\mathrm{Var}(y_2) = \lambda_2$

Geometry of covariance matrix



$\longrightarrow$



$$\tilde{y}_1 = y_1/\sqrt{\lambda_1}$$
$$\tilde{y}_2 = y_2/\sqrt{\lambda_2}$$

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \;=\; (\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{u}})^T(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{u}}) \;=\; ||\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{u}}||^2 \qquad (\dagger)$$

$$\text{where}\quad \tilde{\boldsymbol{u}} = \left(\frac{\boldsymbol{v}_1}{\sqrt{\lambda_1}}, \frac{\boldsymbol{v}_2}{\sqrt{\lambda_2}}\right)^T\boldsymbol{\mu}$$
$$= \left(\frac{\boldsymbol{v}_1^T\boldsymbol{\mu}}{\sqrt{\lambda_1}}, \frac{\boldsymbol{v}_2^T\boldsymbol{\mu}}{\sqrt{\lambda_2}}\right)^T$$

Problems with the estimation of covariance matrix

- $|\boldsymbol{\Sigma}| \rightarrow 0$ when
  - N is not large enough (when compared with D)
    NB: $|\boldsymbol{\Sigma}| = 0$ for $N \le D$
  - There is high dependence (correlation) among variables (e.g. $\rho(x_i, x_j) \approx 1$)
- $\boldsymbol{\Sigma}^{-1}$ becomes unstable when $|\boldsymbol{\Sigma}|$ is small.
- Solutions?
  - Share $\boldsymbol{\Sigma}$ among classes ($\Rightarrow$linear discriminant functions)
  - Assume independence among variables $\Rightarrow$ a diagonal covariance matrix rather than a 'full' covariance matrix.
  - Reduce the dimensionality by transforming the data into a low-dimensional vector space (e.g. PCA).
  - Another regularisation:
    - Add a small positive number to the diagonal elements
      $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} + \epsilon I$

Shared covariance matrix among classes

- How to estimate the shared covariance:
  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all $k = 1, \dots, K$

$$\boldsymbol{\Sigma} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_k$$
$$= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{n=1}^{N_k}(\boldsymbol{x}_n^{(k)}-\boldsymbol{\mu}^{(k)})(\boldsymbol{x}_n^{(k)}-\boldsymbol{\mu}^{(k)})^T$$

- Why is the following not good?

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T$$
$$= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{x}_n^{(k)}-\boldsymbol{\mu})(\boldsymbol{x}_n^{(k)}-\boldsymbol{\mu})^T$$

Covariance matrix when naive Bayes is assumed

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{DD} \end{pmatrix}, \qquad \sigma_{ij} = 0 \text{ for } i \neq j$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$
$$= p(x_1|\mu_1,\sigma_{11})\cdots p(x_D|\mu_D,\sigma_{DD})$$
$$= \prod_{i=1}^{D}\left\{\frac{1}{\sqrt{2\pi\sigma_{ii}}}\exp\left(\frac{-(x_i-\mu_i)^2}{2\sigma_{ii}}\right)\right\}$$

Bayes theorem and probability densities

- Rules for probability densities are similar to those for probabilities:

$$p(x, y) = p(x|y)\,p(y)$$
$$p(x) = \int p(x, y)\,dy$$

- We may mix probabilities of discrete variables and probability densities of continuous variables:

$$p(x, Z) = p(x|Z)\,P(Z)$$

- Bayes' theorem for continuous data x and class C:

$$P(C|x) = \frac{p(x|C)\,P(C)}{p(x)}$$
$$P(C|x) \propto p(x|C)\,P(C)$$

Bayes theorem and univariate Gaussians

- If $p(x|C)$ is Gaussian with mean μ and variance σ^2:

$$P(C|x) \propto p(x|C)\,P(C) \;= N(x;\mu,\sigma^2)\,P(C)$$
$$\propto \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)P(C)$$

- The log likelihood $LL(x|C)$ is:

$$LL(x\,|\,\mu,\sigma^2) = \ln p(x\,|\,\mu,\sigma^2)$$
$$= \frac{1}{2}\left(-\ln(2\pi) - \ln\sigma^2 - \frac{(x-\mu)^2}{\sigma^2}\right)$$

- The log posterior probability $\ln P(C|x)$ is:

$$\ln P(C\,|\,x) \propto LL(x\,|\,C) + \ln P(C)$$
$$\propto \frac{1}{2}\left(-\ln(2\pi) - \ln\sigma^2 - \frac{(x-\mu)^2}{\sigma^2}\right) + \ln P(C)$$

Log probability ratio (log odds)

For a classification problem of two classes: C_1 and C_2,

$$\ln\frac{P(C_1|x)}{P(C_2|x)} = \ln P(C_1|x) - \ln P(C_2|x)$$
$$= -\frac{1}{2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2} + \ln\sigma_1^2 - \ln\sigma_2^2\right)$$
$$+ \ln P(C_1) - \ln P(C_2)$$

$$\ln P(C_1|x) - \ln P(C_2|x) > 0 \quad\Rightarrow\quad C_1$$

$$\ln P(C_1|x) - \ln P(C_2|x) < 0 \quad\Rightarrow\quad C_2$$

Example: 1-dimensional Gaussian classifier

- Two classes, S and T, with some observations:

Class S	10	8	10	10	11	11
Class T	12	9	15	10	13	13

- Assume that each class may be modelled by a Gaussian. The estimated mean and variance of each pdf with the maximum likelihood (ML) estimation are given as follows:

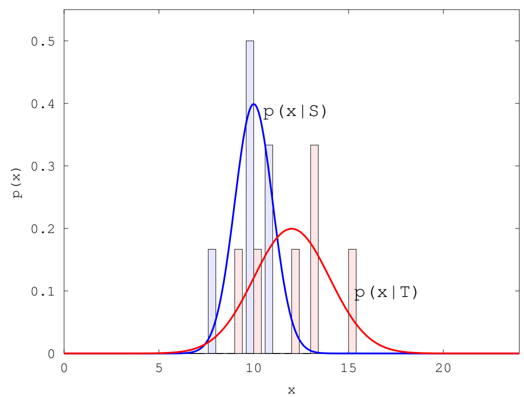
$$\mu(S) = 10 \quad \sigma^2(S) = 1$$
$$\mu(T) = 12 \quad \sigma^2(T) = 4$$

- The following unlabelled data points are available:

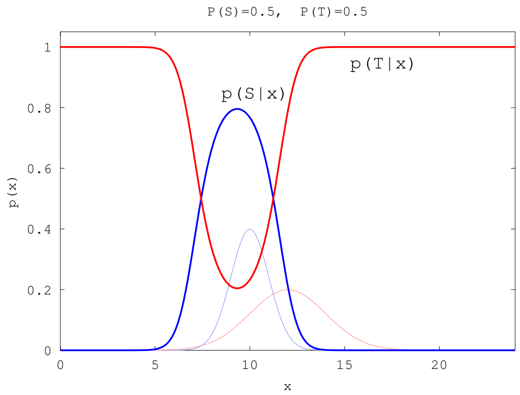
$$x_1 = 10, \quad x_2 = 11, \quad x_3 = 6$$

  To which class should each of the data points be assigned?
  Assume the two classes have equal prior probabilities.

Gaussian pdfs for S and T vs histograms



Posterior probabilities



Example: 1-dimensional Gaussian classifier *(cont.)*

- Take the log odds (posterior probability ratios):

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_s)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln\sigma_S^2 - \ln\sigma_T^2\right) + \ln P(S) - \ln P(T)$$

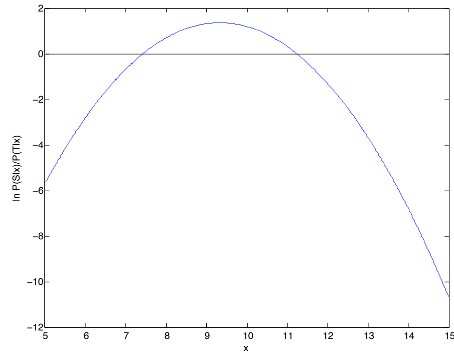
- In the example the priors are equal, so:

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_s)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln\sigma_S^2 - \ln\sigma_T^2\right)$$
$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right)$$

- If log odds are less than 0 assign to T, otherwise assign to S.

Log odds

Test samples: $x_1 = 10,\ x_2 = 11,\ x_3 = 6$



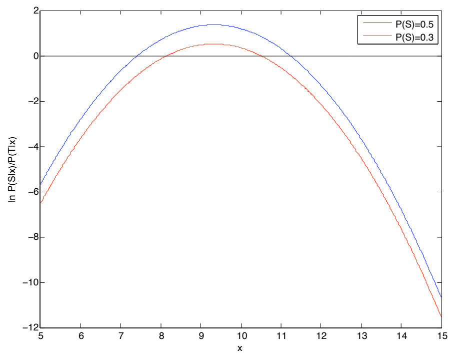
Example: unequal priors

- Now, assume $P(S) = 0.3, P(T) = 0.7$. Including this prior information, to which class should each of the above test data points, x_1, x_2, x_3, be assigned?
- Again compute the log odds:

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_s)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln\sigma_S^2 - \ln\sigma_T^2\right) + \ln P(S) - \ln P(T)$$
$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln P(S) - \ln P(T)$$
$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln(3/7)$$

Log odds

Test samples: $x_1 = 10,\ x_2 = 11,\ x_3 = 6$



Multivariate Gaussian classifier

- Multivariate Gaussian (in D dimensions):

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

- Log likelihood:

$$LL(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

- Posterior probability: $p(C|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\mu},\Sigma)P(C)$
- Log posterior probability:

$$\ln P(C\,|\,\mathbf{x}) \propto -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\Sigma| + \ln P(C) + \text{const.}$$

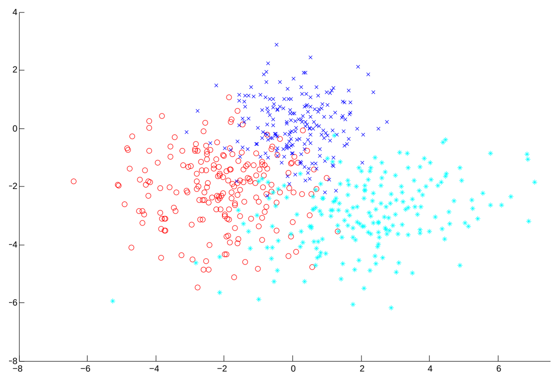
- Try Q4 of Tutorial 4

Example

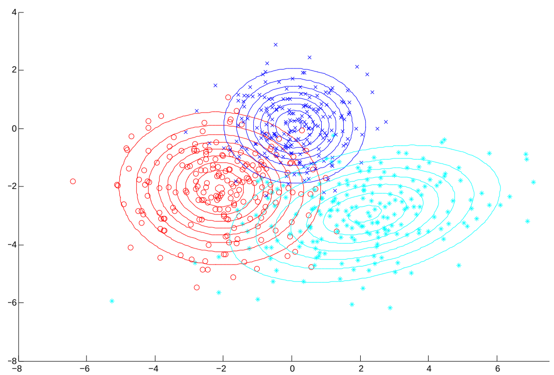
- 2-dimensional data from three classes (A, B, C).
- The classes have equal prior probabilities.
- 200 points in each class
- Load into Matlab ($n \times 2$ matrices, each row is a data point) and display using a scatter plot:

```
xa = load('trainA.dat');
xb = load('trainB.dat');
xc = load('trainC.dat');
hold on;
scatter(xa(:, 1), xa(:,2), 'r', 'o');
scatter(xb(:, 1), xb(:,2), 'b', 'x');
scatter(xc(:, 1), xc(:,2), 'c', '*');
```

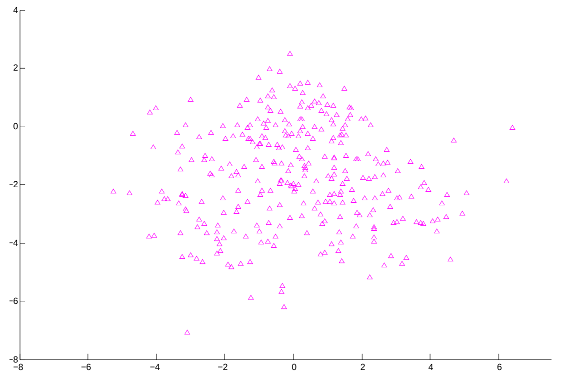

Training data



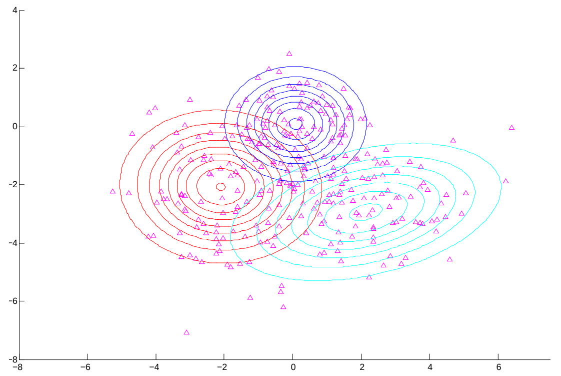
Gaussians estimated from training data



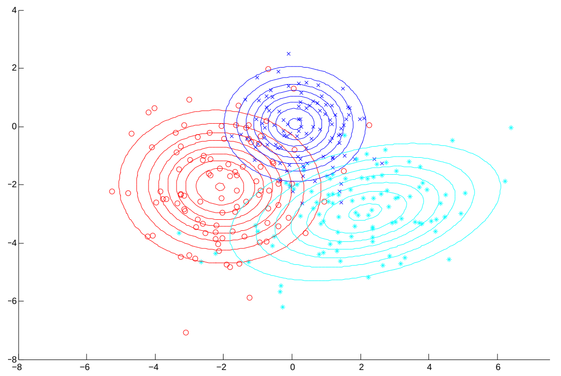
Testing data



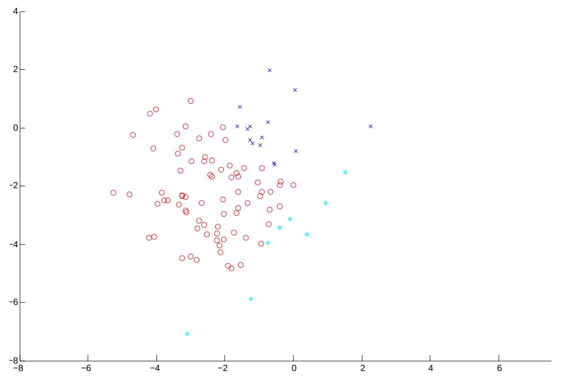
Testing data — with estimated class distributions



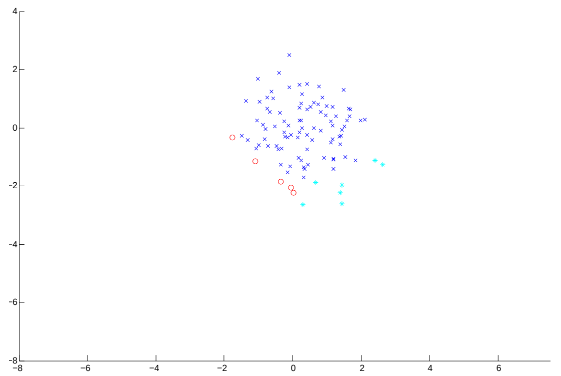
Testing data — with true class indicated



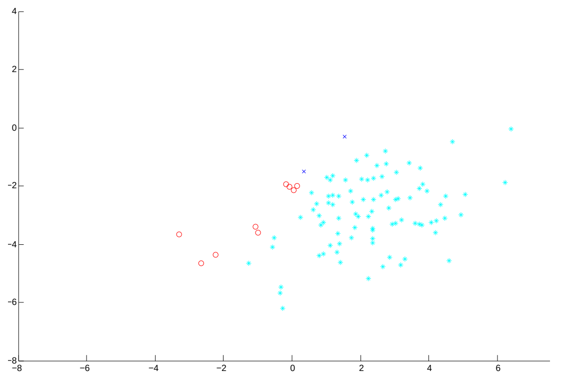
Classifying test data from class A



Classifying test data from class B



Classifying test data from class C



Result

- Analyse the result by percent correct, and in more detail with a confusion matrix
  - Columns of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Rows correspond to the actual (true) class labels
  - Element (r, c) is the number of patterns from true class r that were classified as class c
  - Total number of correctly classified patterns is obtained by summing the numbers on the leading diagonal
- Confusion matrix in this case

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

- Overall proportion of test patterns correctly classified is

$$(77 + 88 + 89)/300 = 254/300 = 0.85$$

Performance measures

- Accuracy (correct classification rate) $= 1 -$ error rate
- Confusion matrix
- Precision, Recall
- F-measure (F1 score)

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Receiver operating characteristic (ROC)

NB: measures shown in grey are non-examinable

Example: Classifying spoken vowels

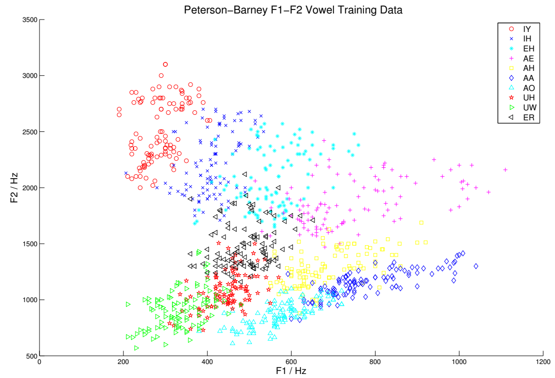
- 10 Spoken vowels in American English
- Vowels can be characterised by formant frequencies — resonances of vocal tract
  - there are usually three or four identifiable formants
  - first two formants written as F1 and F2
- Peterson-Barney data — recordings of spoken vowels by American men, women, and children
  - two examples of each vowel per person
  - for this example, data split into training and test sets
  - children's data not used in this example
  - different speakers in training and test sets
- (see `http://en.wikipedia.org/wiki/Vowel` for more)
- Classify the data using a Gaussian classifier
- Assume equal priors

The data

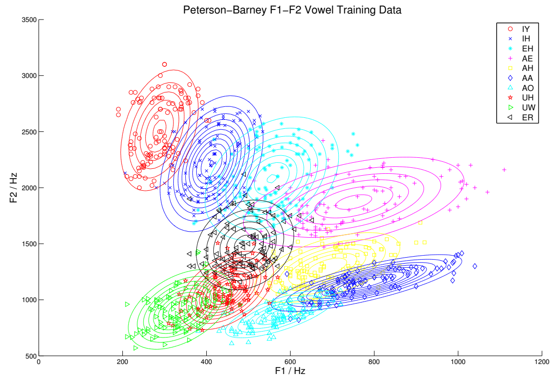
Ten steady-state vowels, frequencies of F1 and F2 at their centre:

- IY — "bee"
- IH — "big"
- EH — "red"
- AE — "at"
- AH — "honey"
- AA — "heart"
- AO — "frost"
- UH — "could"
- UW — "you"
- ER — "bird"

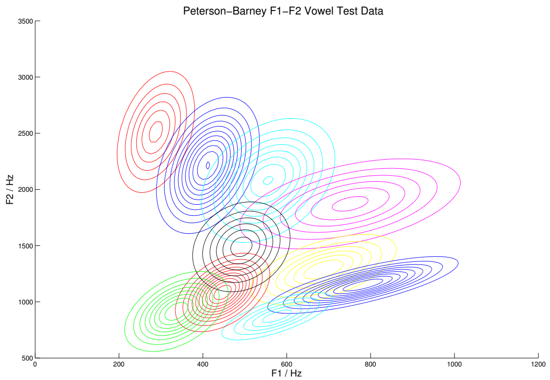
Vowel data — 10 classes



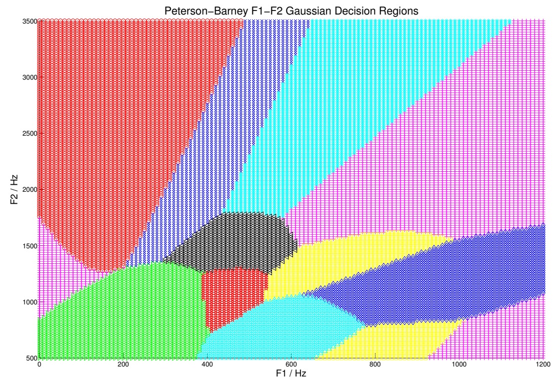
Data and Gaussians for each class



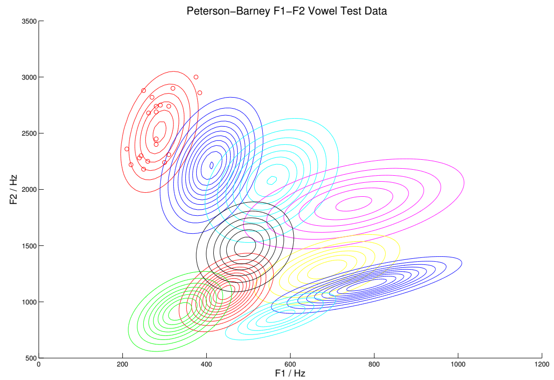
Gaussians for each class



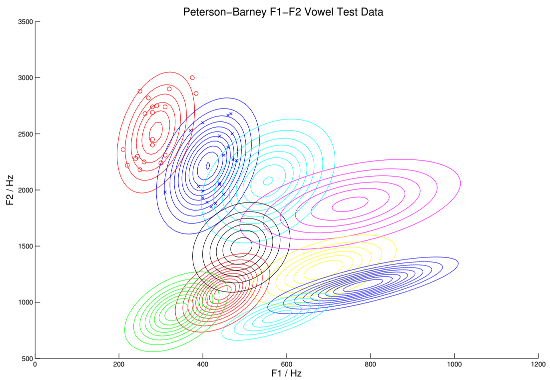
Decision Regions



Test data for class 1 (IY)



Test data for class 2 (IY)



Confusion matrix

	Predicted class										
	IY	IH	EH	AE	AH	AA	AO	UH	UW	ER	% corr.
IY	20	0	0	0	0	0	0	0	0	0	100
IH	0	20	0	0	0	0	0	0	0	0	100
EH	0	0	15	1	0	0	0	0	0	4	75
AE	0	0	3	16	1	0	0	0	0	0	80
AH	0	0	0	0	18	2	0	0	0	0	90
AA	0	0	0	0	2	17	1	0	0	0	85
AO	0	0	0	0	0	4	16	0	0	0	80
UH	0	0	0	0	2	0	0	18	0	0	90
UW	0	0	0	0	0	0	0	5	15	0	75
ER	0	0	0	0	0	0	0	2	0	18	90

Total: 86.5% correct

Exercise

1. Consider estimating a covariance matrix Σ from a data set. Discuss what we could say about the data for the following situations:
   - Σ is almost diagonal (i.e. $\sigma_{ij} \approx 0$ for $i \neq j$).
   - $|\Sigma| \approx 0$.
2. Give examples of data for each situation above.
3. Discuss the minimum number of training samples required to have a covariance matrix that is invertible, i.e. $|\Sigma| \neq 0$. (Hint: think $D = 1$ first, and $D = 2$, and so on.)

Summary

- Covariance matrix
- Using Bayes' theorem with pdfs
- Log probability ratio (log odds)
- The Gaussian classifier: 1-dimensional and multi-dimensional
- Classification examples
- Evaluation measures. Confusion matrix

Familiarise yourself with vector/matrix operations, using pens and papers! (as well as computers)