Inf2b Learning and Data

Lecture 9: Multivariate Gaussians and Classification

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

The multidimensional Gaussian distribution

- The d-dimensional vector $\mathbf{x} = (x_1, \ldots, x_d)^T$ is multivariate Gaussian if it has a probability density function of the following form:
$$p(\mathbf{x}\,|\,\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$
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 $\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is referred to as a *quadratic form*.

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:
$$\boldsymbol{\mu} = E[\mathbf{x}]$$
- The covariance matrix Σ is the expectation of the deviation of $\mathbf{x}$ from the mean:
$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$
- Σ is a $d \times d$ symmetric matrix:
$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \sigma_{ji}\,.$$
- The sign of the covariance helps to determine the relationship between two components:
  - If x_j is large when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be positive;
  - If x_j is small when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be negative.

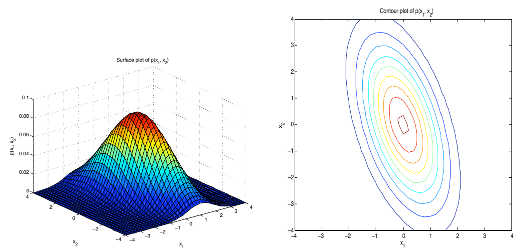
Covariance matrix (cont.)

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \cdots & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \cdots & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \sigma_{ii} & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \cdots & \cdots & \sigma_{dd} \end{pmatrix}$$

- $\sigma_i^2 = \sigma_{ii}$
- $|\Sigma| = \det(\Sigma)$: determinant

e.g. $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = a \times d - b \times c$

2-D Gaussian with a full covariance matrix



$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \qquad \rho_{12} = -0.5$$

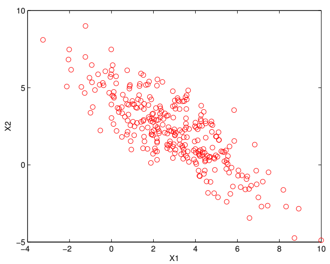
Parameter estimation

Maximum likelihood estimation (MLE):

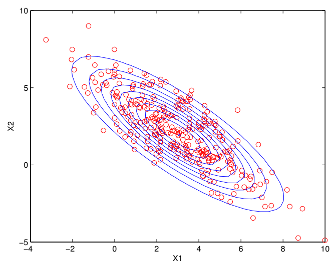
$$\boldsymbol{\mu} = E[\mathbf{x}]$$
$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)}$$

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^T$$

Example data



Maximum likelihood fit to a Gaussian



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_c and variance σ_c^2:

$$P(C|x) \propto p(x|C)\,P(C) \;= N(x; \mu_c, \sigma_c^2)\,P(C)$$

$$\propto \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x-\mu_c)^2}{2\sigma_c^2}\right) P(C)$$

- Taking logs, we have the log likelihood $LL(x|C)$:

$$LL(x\,|\,\mu_c, \sigma_c^2) = \ln p(x\,|\,\mu_c, \sigma_c^2)$$

$$= \frac{1}{2}\left(-\ln(2\pi) - \ln\sigma_c^2 - \frac{(x-\mu_c)^2}{\sigma_c^2}\right)$$

- The log posterior probability $LP(C|x)$ is:

$$LP(C\,|\,x) \propto LL(x\,|\,C) + LP(C)$$

$$\propto \frac{1}{2}\left(-\ln(2\pi) - \ln\sigma_c^2 - \frac{(x-\mu_c)^2}{\sigma_c^2}\right) + \ln P(C)$$

Example: 1-dimensional Gaussian classifier

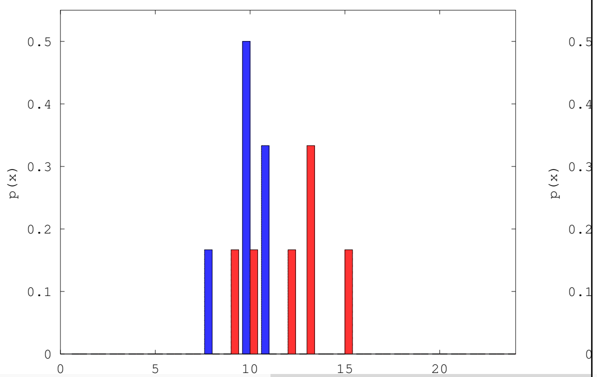
- Two classes, S and T, with some observations:

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 mean and variance of each pdf are estimated by the sample mean and sample variance:

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

Gaussian pdfs for S and T vs histograms



Gaussian pdfs vs histograms

- Parametric methods vs nonparametric methods

- Discuss pros and cons.

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 mean and variance of each pdf are estimated by the sample mean and sample variance:

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

- 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.

Log odds

- Take the log odds (posterior probability ratios):

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

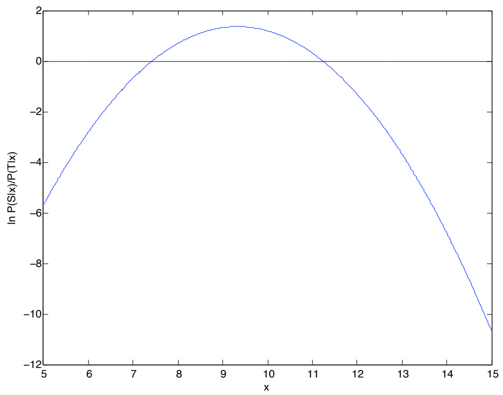
- In the example the priors are equal, so:

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

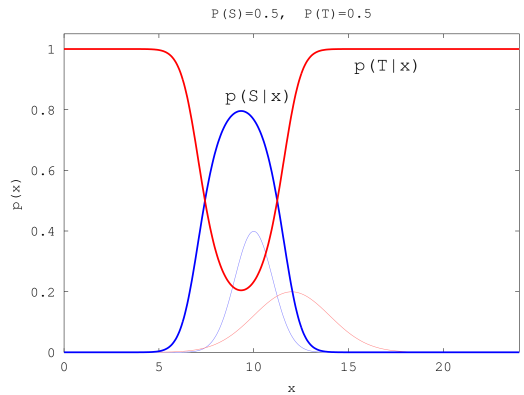
$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right)$$

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

Log odds



Log odds vs Posterior probabilities



Example: unequal priors

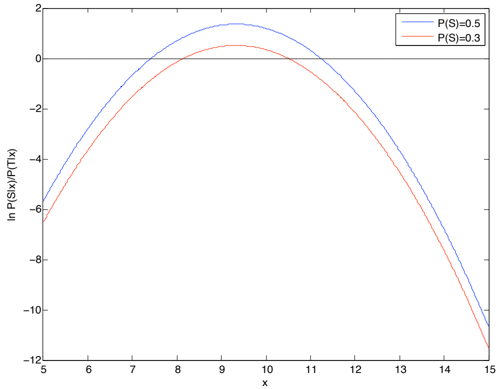
- Now, assume $P(S) = 0.3$, $P(T) = 0.7$. Including this prior information, to which class should each of the above test data points $(x^{(1)}, x^{(2)}, x^{(3)})$ be assigned?
- Again compute the log odds:

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

$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln P(S) - \ln P(T)$$

$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln(3/7)$$

Log odds



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:

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

- If $p(C|\mathbf{x}) \sim p(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, the log posterior probability is:

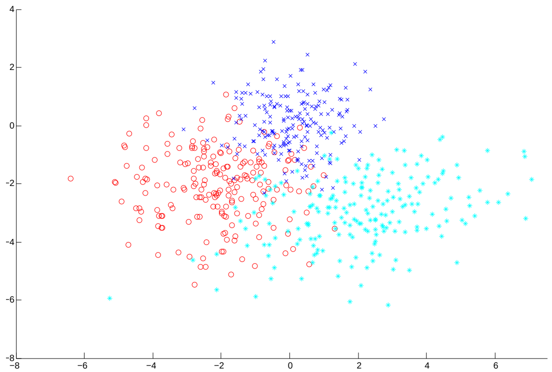
$$\ln P(C\,|\,\mathbf{x}) \propto -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\Sigma| + \ln P(C) + \text{const.}$$

Example

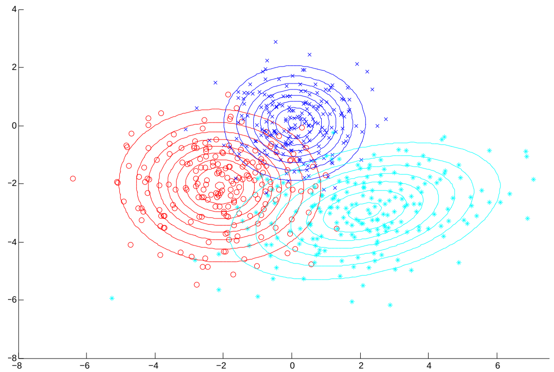
- 2-dimensional data from three classes (A, B, C).
- The classes have equal prior probabilities.
- 200 points in each class
- Load into Matlab ($n \times 2$ matrices, each row is a data point) and display using a scatter plot:

```
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', '*');
```

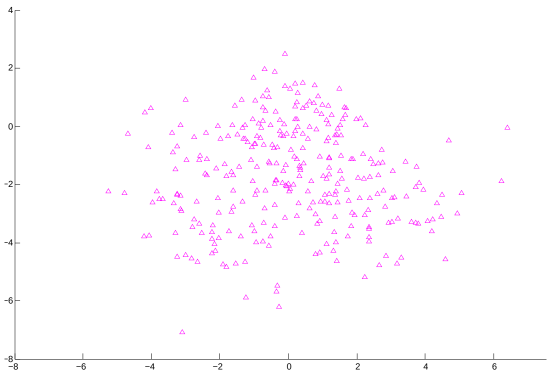

Training data



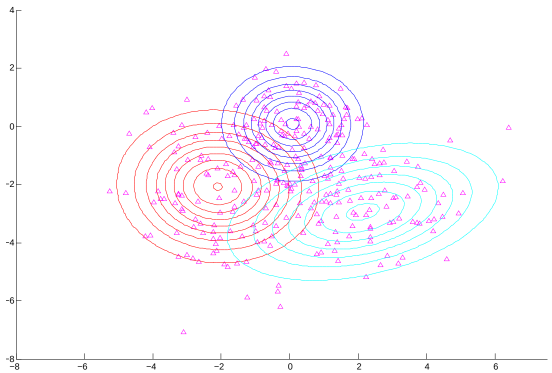
Gaussians estimated from training data



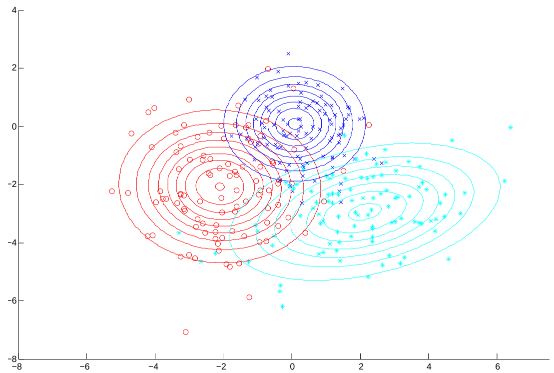
Testing data



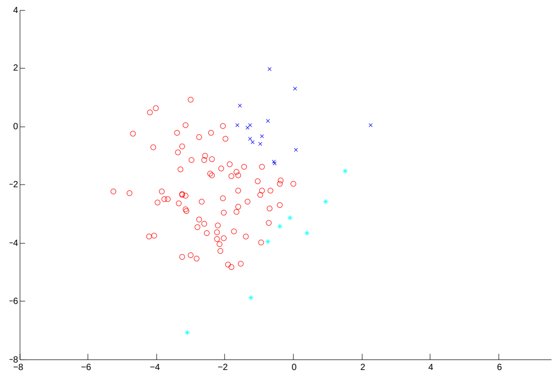
Testing data — with estimated class distributions



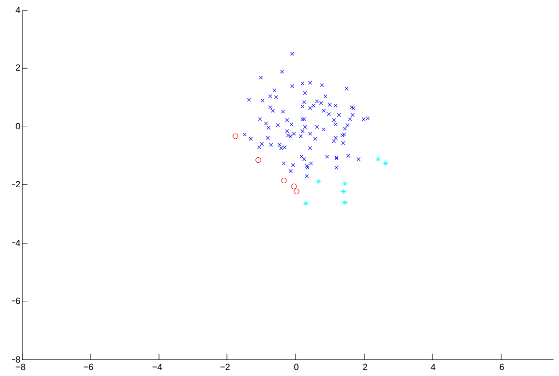
Testing data — with true class indicated



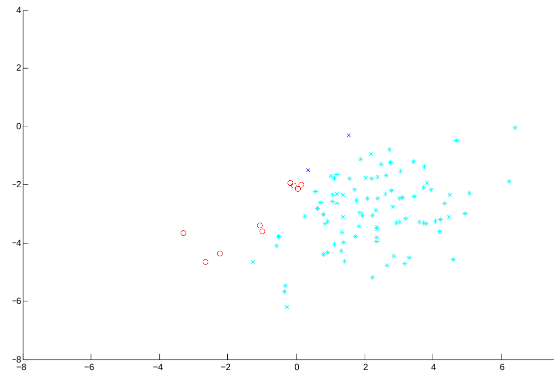
Classifying test data from class A



Classifying test data from class B



Classifying test data from class C



Results

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

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

- Overall proportion of test patterns correctly classified is

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

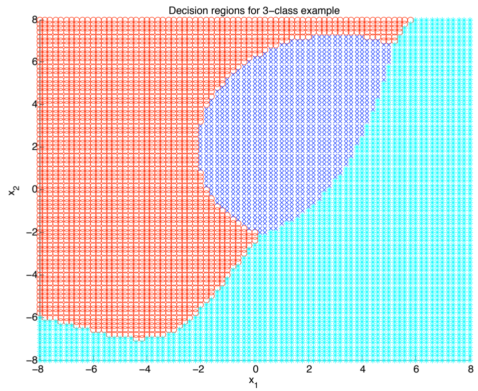
Performance measures

- Precision (correct rate)
- Accuracy
- Confusion matrix
- F-measure (F1 score)

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

- Receiver operating characteristic (ROC)

Decision Regions



Example: Classifying spoken vowels

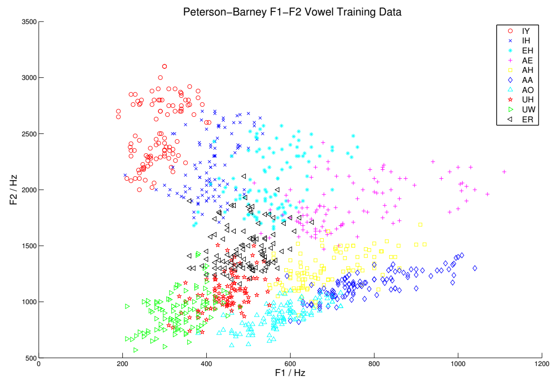
- 10 Spoken vowels in American English
- Vowels can be characterised by formant frequencies — resonances of vocal tract
  - there are usually three or four identifiable formants
  - first two formants written as F1 and F2
- Peterson-Barney data — recordings of spoken vowels by American men, women, and children
  - two examples of each vowel per person
  - for this example, data split into training and test sets
  - childrens 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

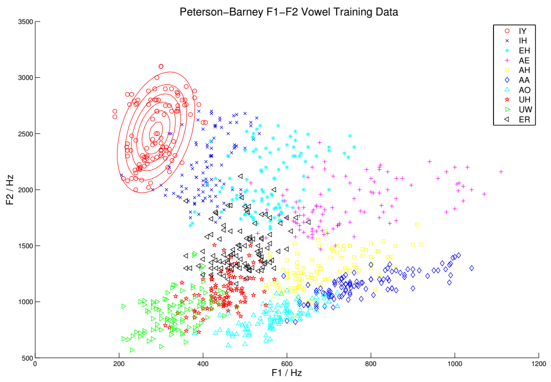
Ten steady-state vowels, frequencies of F1 and F2 at their centre:

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

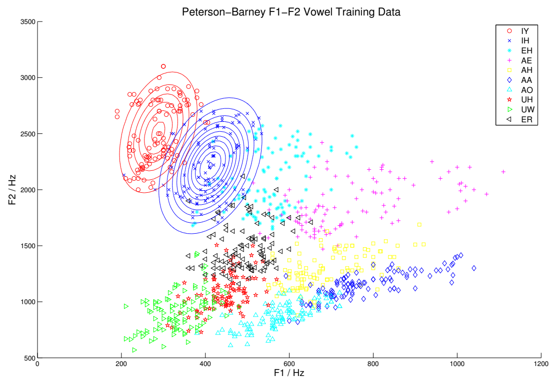
Vowel data — 10 classes



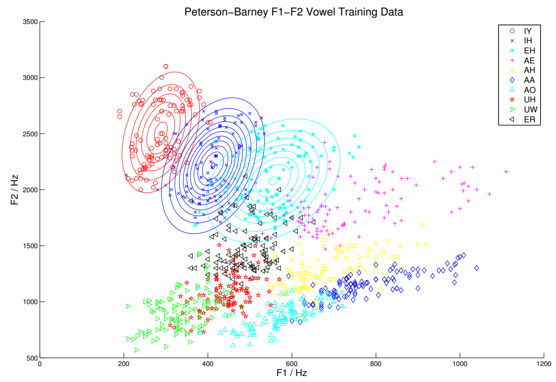
Gaussian for class 1 (IY)



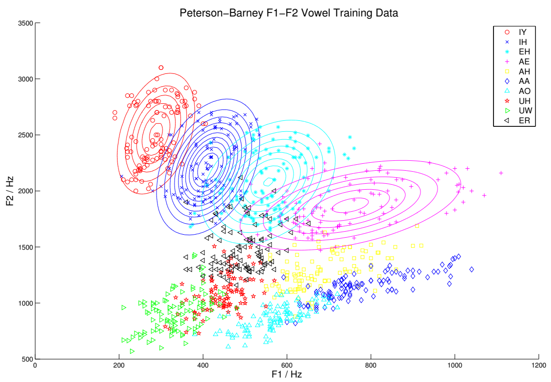
Gaussian for class 2 (IH)



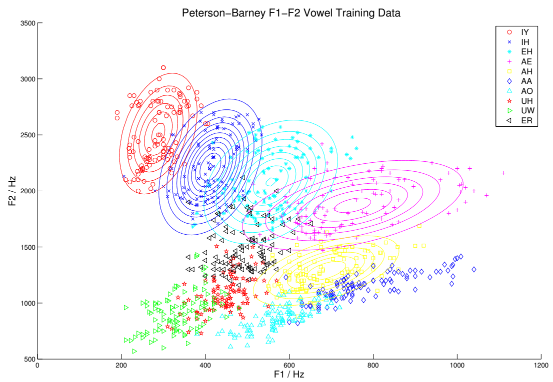
Gaussian for class 3 (EH)



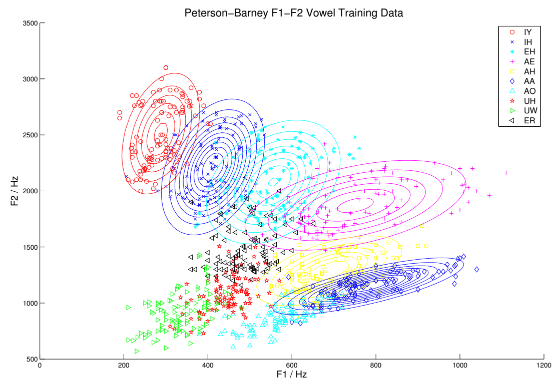
Gaussian for class 4 (AE)



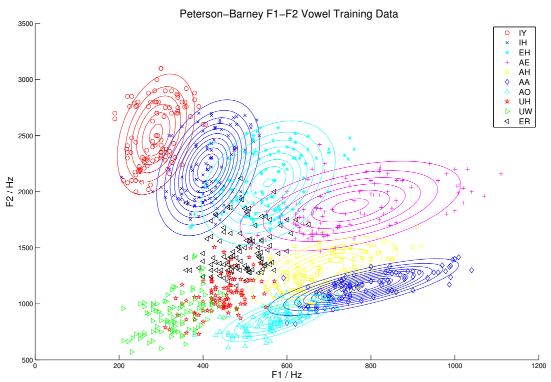
Gaussian for class 5 (AH)



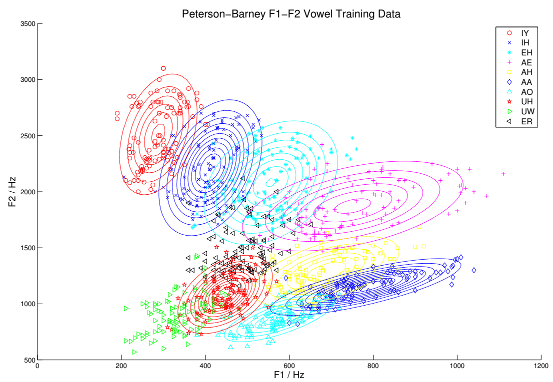
Gaussian for class 6 (AA)



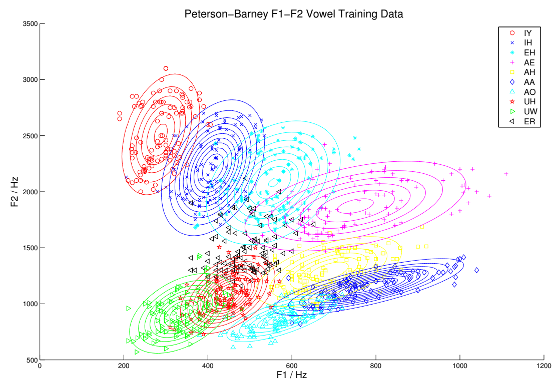
Gaussian for class 7 (AO)



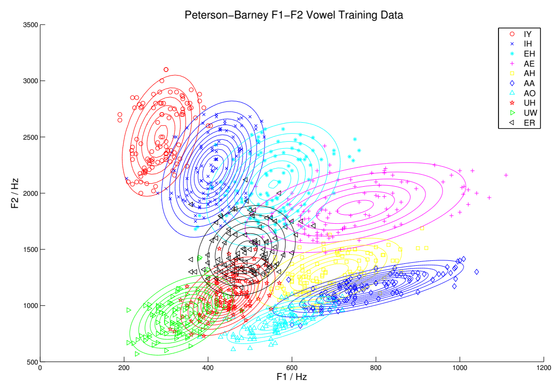
Gaussian for class 8 (UH)



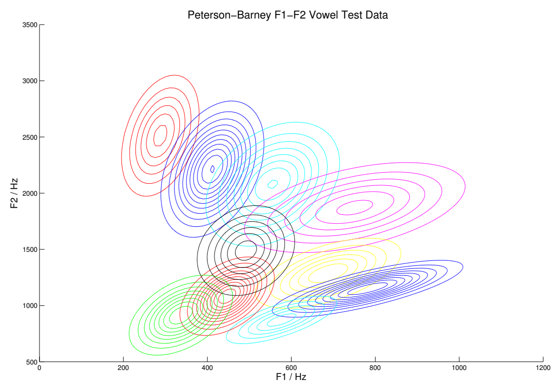
Gaussian for class 9 (UW)



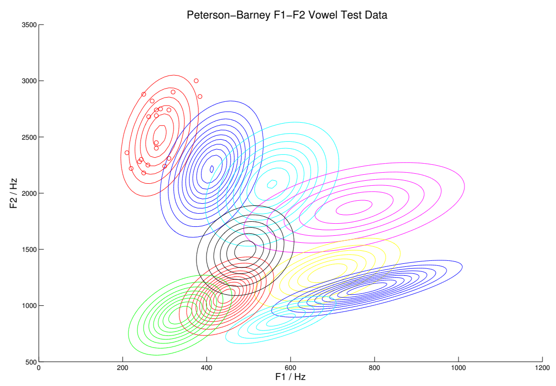
Gaussian for class 10 (ER)



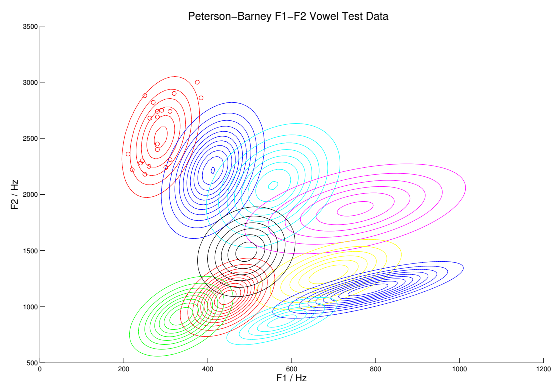
Gaussians for each class)



Test data for class 1 (IY))



Test data for class 2 (IY))



Final confusion matrix

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

Total: 86.5% correct

Summary

- Using Bayes' theorem with pdfs
- The Gaussian classifier: 1-dimensional and multi-dimensional
- Classification examples
- Confusion matrix