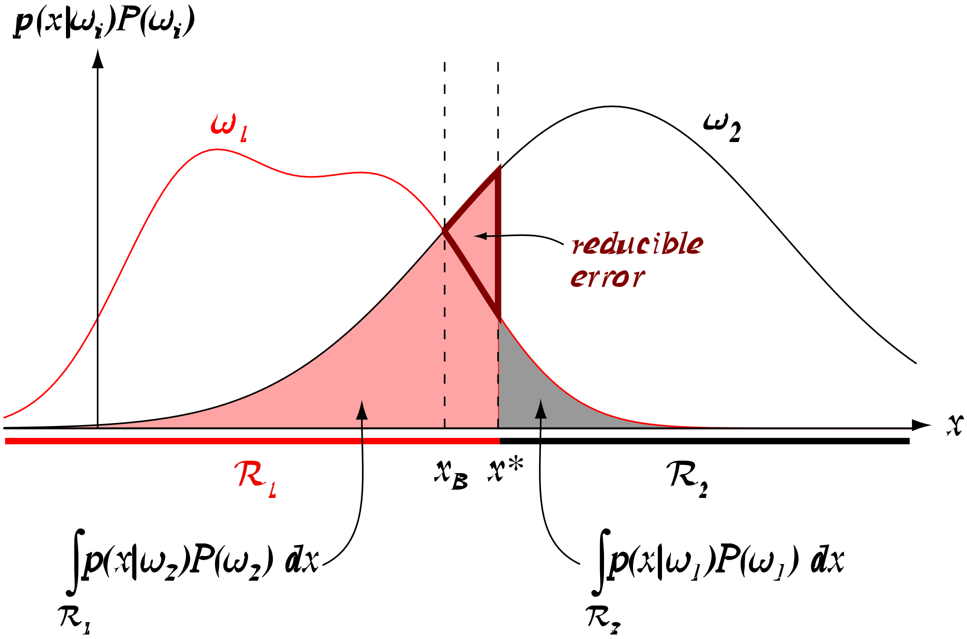
University of Cambridge
Engineering Part IIB

Module 4F10: Statistical Pattern Processing

Handout 1: Introduction & Decision Rules



Mark Gales
mjfg@eng.cam.ac.uk

Michaelmas 2015

Syllabus

1. Introduction & Bayes' Decision Theory (1L)

- Statistical pattern processing
- Bayesian decision theory
- Classification cost & ROC curves

2. Multivariate Gaussians & Decision Boundaries (1L)

- Decision boundaries for Multivariate Gaussians
- Maximum likelihood estimation

3. Gaussian Mixture Models (1L)

- Mixture models
- Parameter estimation
- EM for discrete random variables

4. Expectation Maximisation (1L)

- Latent variables both continuous and discrete
- Proof of EM

5. Mixture and Product of Experts (1L)

- Gating functions
- Mixtures versus Product of Experts
- Product of Gaussian experts

6. Restricted Boltzman Machines (1L)

- RBM structure
- Contrastive divergence

Syllabus (cont)

5. Linear Classifiers (1L)

- Single layer perceptron
- Perceptron learning algorithm

6. Multi-Layer Perceptrons (2L)

- Basic structure
- Gradient descent parameter optimisation
- Deep topolgies and network initialisation

7. Support Vector Machines (2L)

- Maximum margin classifiers
- Training SVMs
- Kernel functions & Non-linear SVMs

9. Classification and Regression Trees (1L)

- Decision trees
- Query selection
- Multivariate decision trees

10. Non-Parametric Techniques (1L)

- Parzen windows
- Nearest neighbour rule
- K-nearest neighbours

11. Application: Speaker Verification/Identification (1L)

- Speaker recognition/verification task
- GMMs and MAP adaptation
- SVM-based verification

Overview of Course

Generative Models

- Multivariate Gaussian
- Gaussian Mixture Model
- RBMs

Discriminative Classifiers

Discriminative Functions

- Perceptron Algorithm
- Support Vector Machine

Discriminative Models

- Logistic Classification
- Multilayer Perceptron

Non−Parametric

- Decision Trees
- Parzen Windows
- Nearest Neighbour

Training Criteria

- Maximum Likelihood
- Expectation Maximisation
- Maximum Margin

Course Structure

Total of 14L + 2 Examples Classes

Lecturer: Mark Gales

Web-Page: http://mi.eng.cam.ac.uk/~mjfg/4F10/index.html

Assessment by exam (1.5h): 3 questions from 5.

A number of books cover parts of the course material.

- C.M.Bishop, *Pattern Recognition and Neural Networks* OUP, 1995, `CUED: NOF 55`
- R.O.Duda, P.E.Hart & D.G. Stork *Pattern Classification*, Wiley, 2001, `CUED: NOF 64`
- D.J.C. Mackay, *Information Theory, Inference and Learning Algorithms*, CUP, 2004. `CUED: NO 277`
- C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006.

Statistical Pattern Processing

> In this world nothing can be said to be certain, except death and taxes.

- Benjamin Franklin

We make decisions under uncertainty all the time

- gambling (not recommended)
- weather forecasting (not very successfully)
- insurance (risk assessment)
- stock market

Need to formalise "intuitive decisions" mathematically

Basically, how to quantify and manipulate uncertainty.

This course will concentrate on classification, however regression and clustering will be briefly mentioned.

A range of statistical approaches to decision making will be examined:

- approaches can be trained (hence Machine Learning)
- wide range of applications, examples are ...

Automatic Speech Recognition

- Increasing number of speech applications
- Siri: “intelligent” personal assistant
  - speech recognition/synthesis
  - limited “understanding”
- Google: Talk, Voice Search
  - generate texts
  - mobile-phone spoken queries
- Xbox: Kinect interface
  - game console control

Information Retrieval

viavoice image - Google Search http://www.google.co.uk/search?q=viavoice+im...

Sign in

Google Web Images News Maps New! Products Groups

viavoice image

Search: the web pages from the UK New! View and manage your web history

Web Results **1** - **10** of about **315,000** for **viavoice image**. (**0.23** seconds)

IBM **ViaVoice** 10 Sponsored Link
www.Nuance.co.uk Dictate, edit and correct text with your **voice**. Official site.

Image: IBM **ViaVoice** 10.0 Standard Edition: ScanSoft
Image: IBM **ViaVoice** 10.0 Standard Edition: ScanSoft by ScanSoft.
www.amazon.co.uk/ IBM-**ViaVoice**-10-0-Standard-Edition/dp/**image**s/B00006IIF2 - 31k - Cached - Similar pages

Image: IBM **ViaVoice** 10.0 Pro USB Edition with Headset: ScanSoft
Image: IBM **ViaVoice** 10.0 Pro USB Edition with Headset: ScanSoft by ScanSoft.
www.amazon.co.uk/ IBM-**ViaVoice**-10-0-Pro-Headset/dp/**image**s/B00006IIF1 - 31k - Cached - Similar pages
[More results from www.amazon.co.uk]

Image: IBM **VIAVOICE** Advanced 10.0
Image: IBM **VIAVOICE** Advanced 10.0. ... IBM **VIAVOICE** Advanced 10.0. Close window.
www.amazon.ca/ H009A-G00-10-0-IBM-**VIAVOICE**-Advanced-10-0/dp/**image**s/B0000A58IW - 29k - Cached - Similar pages

Image: IBM **ViaVoice** Standard v.10
Image: IBM **ViaVoice** Standard v.10. ... IBM **ViaVoice** Standard v.10. Close window.
www.amazon.ca/ H109A-G00-10-0-IBM-**ViaVoice**-Standard-v-10/dp/**image**s/B0000A58IV - 31k - Cached - Similar pages
[More results from www.amazon.ca]

Image: IBM **ViaVoice** Pro USB Edition
Image: IBM **ViaVoice** Pro USB Edition. ... IBM **ViaVoice** Pro USB Edition.
www.amazon.com/IBM-**ViaVoice**-Pro-USB-Edition/dp/**image**s/B0000A58IX - 32k - Cached - Similar pages

Sponsored Links

Picture
Picasa creates amazing pictures! See demo, download free from Google
picasa.google.co.uk

Viavoice
Save On **Viavoice**
Fast Shipping, Order Online Now.
www.dabs.com

viavoice
Buy It Cheap On eBay
Low Prices, New and Used
ebay.co.uk

Viavoice at Amazon.co.uk
Get your Software at Amazon.co.uk
Free Delivery on orders over £15
www.amazon.co.uk/software

Viavoice
Compare Prices on Software! Great Deals - Buy & Save Online Today.
www.kelkoo.co.uk/Software

1 of 2 17/09/07 14:54

- Search engines are essential
  - query to find previous image viavoice image
  - determine which pages to return
  - what adverts to include (Google needs to make money)

Statistical Machine Translation

Rafales de marque - lecteur dans la technologie de... http://66.249.91.104/translate_c?hl=en&langpai...

UNIVERSITY OF CAMBRIDGE — DEPARTMENT OF ENGINEERING

Marquer les rafales

Les rafales de marque est un lecteur dans la technologie de l'information dans le laboratoire d'intelligence de machine (autrefois le groupe de vision et de robotique de la parole (SVR)) et un camarade de l'université d'Emmanuel. Il est un membre du groupe de recherche de la parole ainsi que les jeunes de Steve de membres de personnel de corps enseignant, la régfion boisée et la facture Byrne de Phil.

Une brève biographie est accessible en ligne.

[Recherche | projets | publications | étudiants | enseignant | contact]

Intérêts de recherches

- Reconnaissance de la parole continue de grand vocabulaire
- Reconnaissance de la parole robuste
- Adaptation d'orateur
- Étude de machine (en particulier choix modèle et méthodes grain-basées)
- Identification et vérification d'orateur

Une brève introduction à la reconnaissance de la parole est accessible en ligne.
dessus

Projets de recherche

Projets en cours :

- Bruit ASR robuste (Europe Ltd de recherches de Toshiba placée)
- Traitement averti d'environnement rapide et robuste (Europe Ltd de recherches de Toshiba placée)
  - NEW Position d'associé de recherches disponible
- AGILE (projet placé par GALE de DARPA)
- Version 3 de HTK - HTK_V3.4 et exemples sont disponibles.

Projets récemment réalisés :

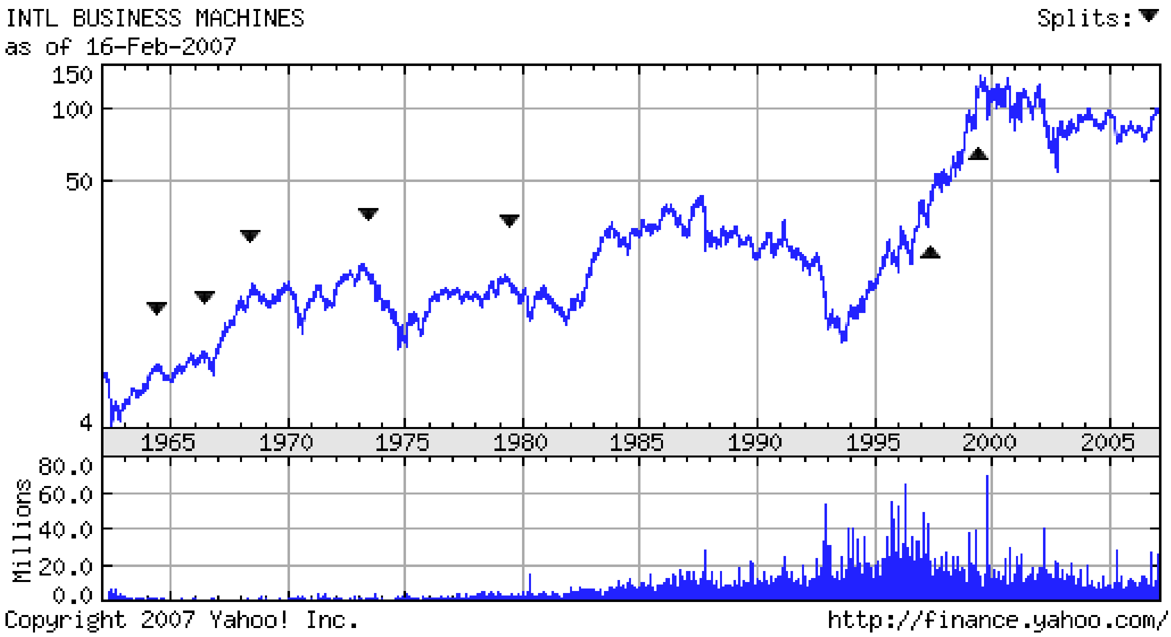
- CoreTex (améliorant la technologie de reconnaissance de la parole de noyau)
- Transcription audio riche de HTK(Projet placé par OREILLES de DARPA) - pages Web locaux

dessus

1 of 3 17/09/07 15:08

- Automatic translation of my web-page using Google in 2007
  - Mark Gales becomes To mark the gusts
  - not quite right yet (but fixed in 2009 version)!

Stock Market Prediction



- Would like to make predictions and actions
- Should I sell my IBM shares?
  - Oct 2007 students voted that I should sell at $100
  - Oct 2008 IBM was at $116.96
  - Oct 2009 IBM was at $120.82
  - Oct 2010 IBM was at $135.48
  - Oct 2011 IBM was at $179.17
  - Oct 2013 IBM was at $184.96
  - Oct 2014 IBM is at $187.17

    though not a linear rise!

What is Statistical Pattern Processing?

The main area of Statistical Pattern Processing discussed in this course is **classification** of patterns into different classes. These patterns can represent many different types of object (speech/images/text etc).

A key issue in all pattern recognition systems is variability. Patterns arise (often from natural sources) that contain variations.

Key issue:

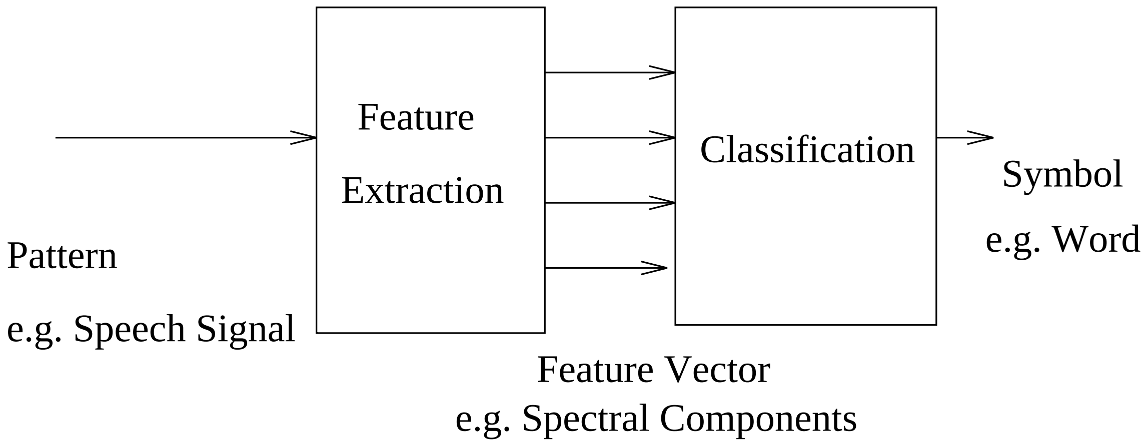
- are the variations systematic (and can be used to distinguish between classes)
- or are they noise

The variability of classes will be approached by using probabilistic modelling of pattern variations.

The standard model for pattern recognition divides the problem into two parts:

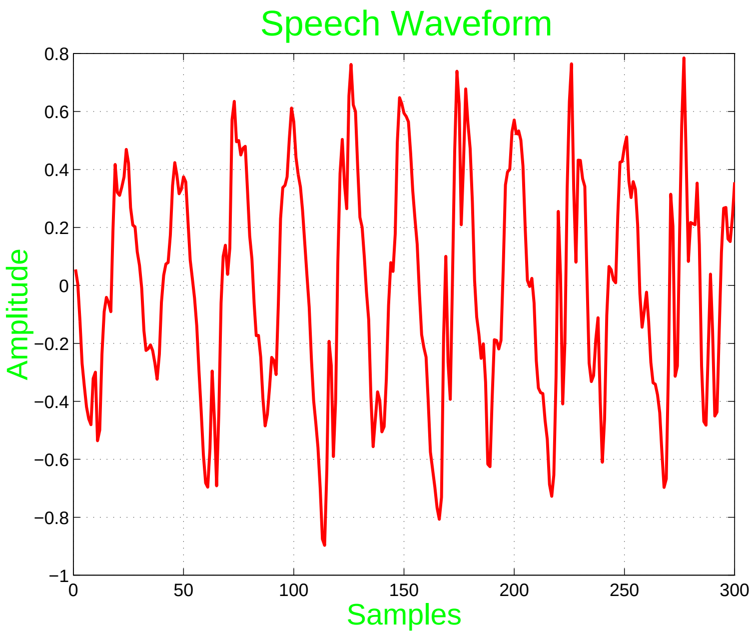
- feature extraction
- classification

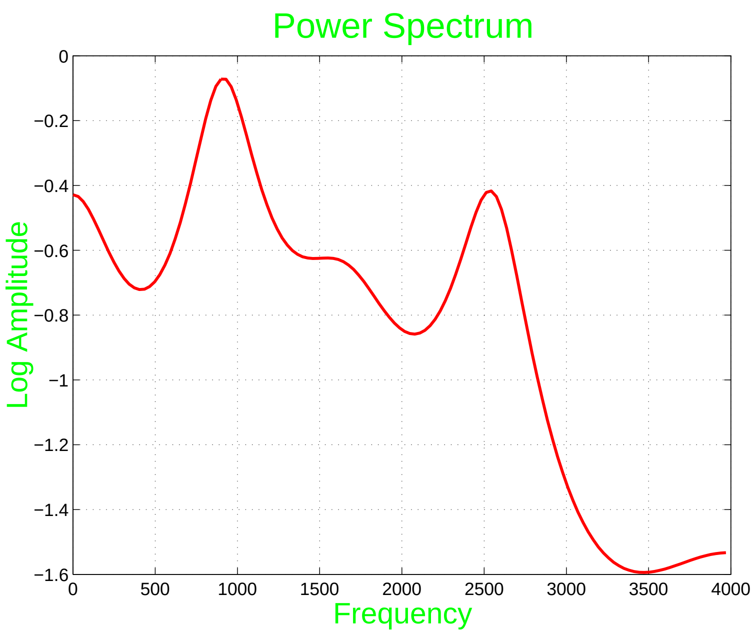
Basic Model



- Initial feature extraction produces a vector of features that contain all the information for subsequent processing (such as classification).
- Ideally, for classification, only the features that contain discriminatory information are used.
- Often features to measure are determined by an “expert”, although techniques exist for choosing suitable features.
- The classifier processes the vector of features and chooses a particular class.
- Normally the classifier is “trained” using a set of data for which there are labelled pairs of feature vectors / class identifiers available.

Simple Speech Features

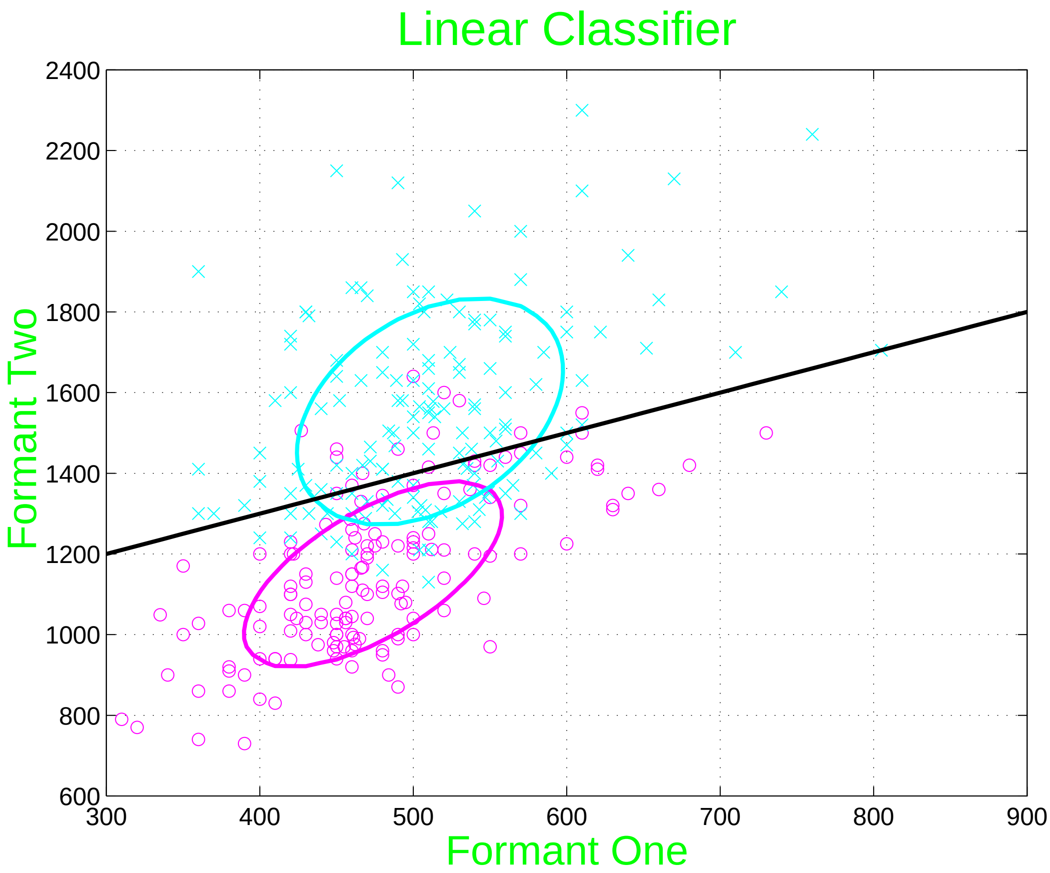




- Features for vowel classification may be the spectral shape or frequencies of peaks (formants)

Simple Vowel Classifier

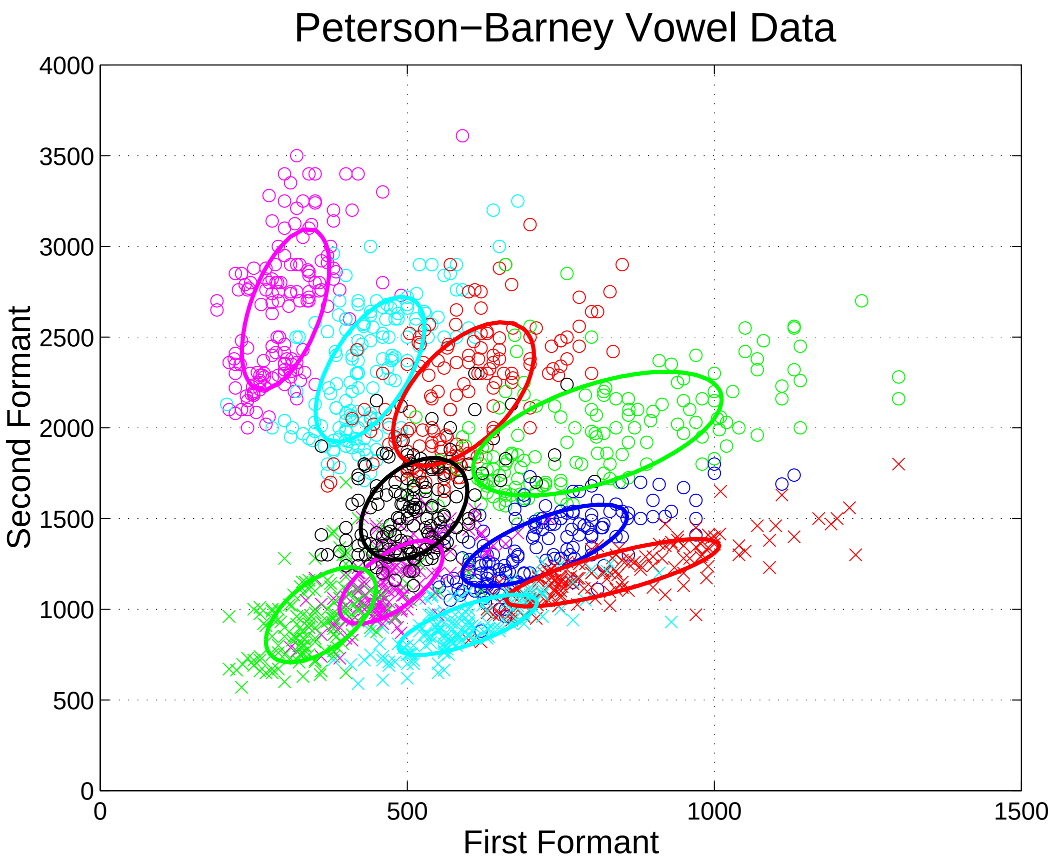
Select two vowels to classify with a linear decision boundary



Most of the data is correctly classified but classification errors occur

- pronunciation of vowels vary from speaker to speaker
- pronunciation vary for a speaker as well!

Vowel Distributions Using Formants



- Vowel classes are reasonably separated (but some overlap!) using these features: could draw decision boundaries
- It is often useful to calculate the probability of a particular class (rarely is this 1/0!)

Some Basic Probability (Revision!!)

- Discrete random variable x takes one value from the set

  $$\mathcal{X} = \omega_1, \ldots, \omega_K$$

  We can compute a set of probabilities

  $$p_j = \Pr(x = \omega_j), \quad j = 1, \ldots, K$$

  We use a probability mass function $P(x)$, to describe the set of probabilities. The PMF satisfies

  $$\sum_{x \in \mathcal{X}} P(x) = 1, \quad P(x) \geq 0$$

- Continuous random variable: scalar x or a vector $\mathbf{x}$. Described by its probability density function (PDF), $p(x)$. The PDF satisfies

  $$\int_{-\infty}^{\infty} p(x) dx = 1, \quad p(x) \geq 0$$

- For random variables x, y, z need

  conditional distribution: $p\left(x|y\right) = \frac{p(x,y)}{p(y)}$

  joint distribution $p\left(x,\, y\right)$

  marginal distribution $\quad p\left(x\right) = \int_{-\infty}^{\infty} p(x,y) dy$

  chain rule $p(x,y,z) = p(x|y,z)\, p(y|z)\, p(z)$

Forms of Classifiers

General notation used in this course

- Observations: each observation consists of a d-dimensional feature vector, $\boldsymbol{x}$.
- Classes (labels): each observation will belong to a single class, $\omega_1, \ldots, \omega_K$

We need a classifier that given an observation, $\boldsymbol{x}$, correctly assigns it to a class, ω.

Classifiers can be split into three broad classes. In the first two a mapping from observation to class can be inferred (the decision rule), the third directly estimates a mapping.

- Generative models: a model of the joint distribution of observations and classes is trained, $p(\boldsymbol{x}, \omega)$.
- Discriminative models: a model of the posterior distribution of the class given the observation is trained , $P(\omega|\boldsymbol{x})$.
- Discriminant functions: a mapping from an observation $\boldsymbol{x}$ to a class ω is directly trained. No posterior probability, $P(\omega|\boldsymbol{x})$, generated just the class label.

See Bishop for a discussion of the merits of these.

Forms of training

Irrespective of the form of classifier the classifier will need to be trained. There are three basic forms of training:

- Supervised learning: for each of the observations, $\boldsymbol{x}$, the correct class label, ω, is available
- Unsupervised learning: only the observation, $\boldsymbol{x}$, is available
- Reinforcement learning: a set of rewards are associated with actions for each observation

This course concentrates on supervised learning.

The training data occurs in pairs. For a 2-class, binary, problem the training data would be

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$$

where

$$y_i = \begin{cases} \omega_1 & \text{if } \boldsymbol{x}_i \text{ generate by class 1} \\ \omega_2 & \text{if } \boldsymbol{x}_i \text{ generate by class 2} \end{cases}$$

The total number of samples from class 1 will be labelled n_1 and for class 2 n_2.

Decision Rules

A "sensible" approach to design a decision rules for generative and discriminative models is to minimises the probability of error:

$$\begin{aligned} P(\text{error}) &= \int P(\text{error}, \boldsymbol{x}) d\boldsymbol{x} \\ &= \int P(\text{error}|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

For a two class problem, the conditional probability of error, (*i.e.* the error probability, given a value for the feature vector), can be written as

$$P(\text{error}|\boldsymbol{x}) = \begin{cases} P(\omega_1|\boldsymbol{x}) & \text{if we decide } \omega_2 \\ P(\omega_2|\boldsymbol{x}) & \text{if we decide } \omega_1 \end{cases}$$

A decision rule that can minimise this conditional probability of error averaged over all samples is required. This leads to Bayes' decision rule, which for a two class problem is

$$\text{Decide} \begin{cases} \text{Class } \omega_1 & \text{if } P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}); \\ \text{Class } \omega_2 & \text{Otherwise} \end{cases}$$

Applying Bayes' decision rule to multi-classes yields

$$\text{Decide } \underset{\omega_j}{\text{argmax}} \{P(\omega_j|\boldsymbol{x})\}$$

Generative Models

For generative models the joint distribution is estimated. For Bayes' decision rule the class posterior is required - this can be obtained using Bayes' rule

$$P(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \omega_j)}{\sum_i p(\boldsymbol{x}, \omega_i)} = \frac{p(\boldsymbol{x}|\omega_j)P(\omega_j)}{p(\boldsymbol{x})}$$

Bayes' rule here computes the posterior probability of a particular class, $P(\omega_j|\boldsymbol{x})$ using the

- likelihood of the data from the class conditional density $p(\boldsymbol{x}|\omega_j)$.
- prior probability of the class ω_j, $P(\omega_j)$ - this is the probability of the class before any data is observed.

The denominator, $p(\boldsymbol{x})$, is sometimes termed the evidence and is the probability density of the data independent of class.

Bayes' Rule is sometimes remembered as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Why Generative Models?

Modelling the joint distribution $p(\boldsymbol{x}, \omega)$ is more complicated than estimating the posterior or a decision boundary.

However classification tasks such as speech recognition commonly use generative models!

Why use generative models all?

- Prior distributions easy to interprete. Estimating the class priors is normally performed by simply taking ML estimate from the counts e.g. $n_1/(n_1 + n_2)$.
- Class-conditional PDF easy to interprete. For speech recognition we can extract the portions of speech associated with a particular word and find the "best" model for that segment.
- Parameters easy to interprete. For speech recognition easy to consider how to adapt the model parameters to a particular speaker.

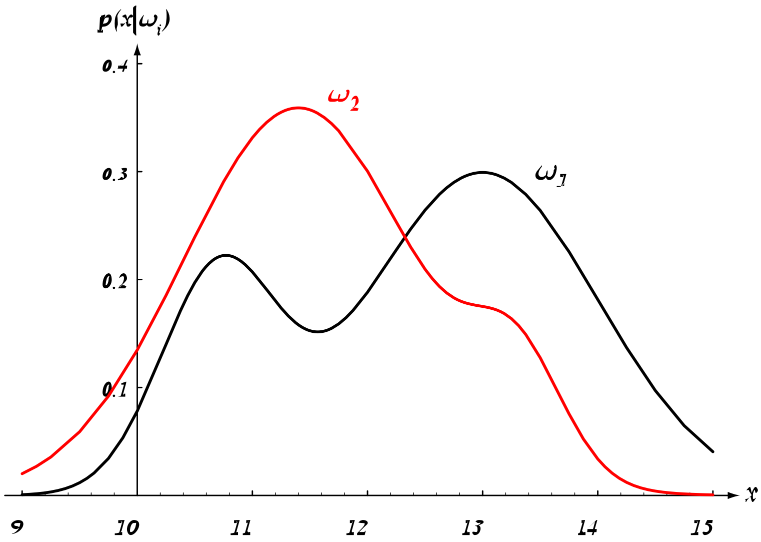
Example

The figures below (from DHS) give hypothetical class-conditional pdfs for two classes:

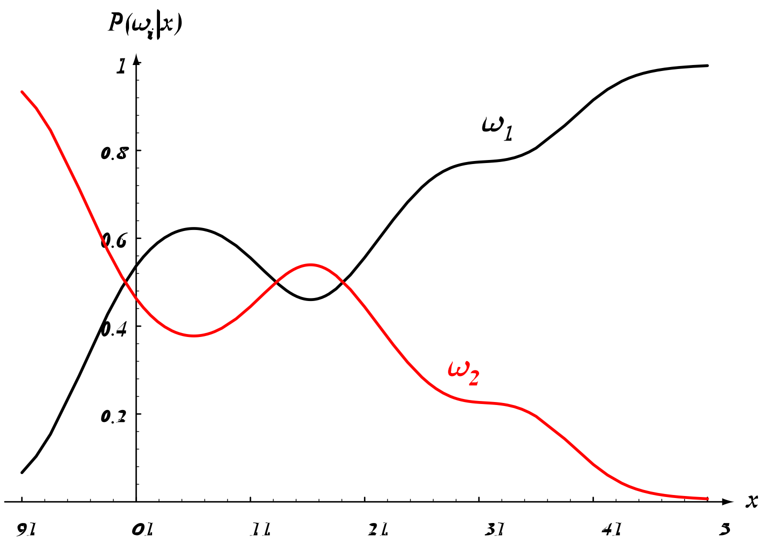
- priors are $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$,

and the posterior distribution.

PDFs:



Posteriors:



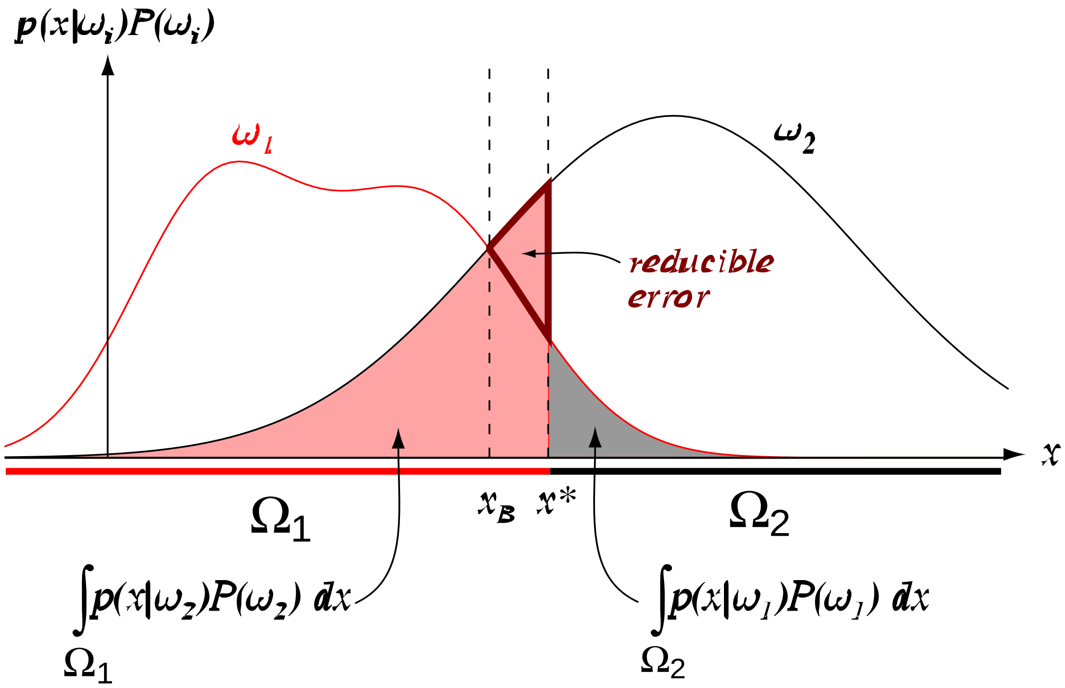
Probability of Error

For a 2-class problem the decision rule will split the observation space into two regions

- Ω_1: observation classified as ω_1
- Ω_2: observation classified as ω_2

$$\begin{aligned} P(\text{error}) &= P(\boldsymbol{x} \in \Omega_2, \omega_1) + P(\boldsymbol{x} \in \Omega_1, \omega_2) \\ &= P(\boldsymbol{x} \in \Omega_2|\omega_1)P(\omega_1) + P(\boldsymbol{x} \in \Omega_1|\omega_2)P(\omega_2) \\ &= \int_{\Omega_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} + \int_{\Omega_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x} \end{aligned}$$

The error regions for a two-class problem are shown below (from DHS). The decision boundary x^* is set to x_B for minimum error.



Generative Model Decision Rule

For the two-class case the Bayes' minimum decision rule can be written as

$$\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1, \qquad \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

The first is the ratio of the posteriors, the second the ratio of the likelihoods compared to the ratio of the priors.

For multi-class problems, the posteriors of the K classes can be calculated

$$P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x}), \ldots, P(\omega_K|\boldsymbol{x})$$

and the largest selected, or use

$$\text{Decide } \underset{\omega_j}{\operatorname{argmax}} \{p(\boldsymbol{x}|\omega_j)P(\omega_j)\}$$

since the RHS denominator of Bayes' rule is independent of class and this is a frequent statement of Bayes' decision rule for minimum error with generative models.

Cost of Mis-Classification

So far the decision rule has aimed to minimise the average probability of classification error. Recall that for the two-class problem, the Bayes minimum average error decision rule can be written as:

$$\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} 1$$

Sometimes, the cost (or loss) for misclassification is specified (or can be estimated) and different types of classification error may not have equal cost.

C_{12}	Cost of choosing $\omega_1|\boldsymbol{x}$ from ω_2
C_{21}	Cost of choosing $\omega_2|\boldsymbol{x}$ from ω_1

and C_{ii} is the cost of correct classification.

The aim now is to minimise the Bayes' Risk which is the expected value of the classification cost.

Minimum Bayes' Risk

Again let the decision region associated with class ω_j be denoted Ω_j. Consider all the patterns that belong to class ω_1. The expected cost (or risk) for these patterns $\mathcal{R}_1$ is given by

$$\mathcal{R}_1 = \sum_{i=1}^{2} C_{i1} \int_{\Omega_i} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x}$$

The overall cost $\mathcal{R}$ is found as

$$\begin{aligned}
\mathcal{R} &= \sum_{j=1}^{2} \mathcal{R}_j P(\omega_j) \\
&= \sum_{j=1}^{2} \sum_{i=1}^{2} C_{ij} \int_{\Omega_i} p(\boldsymbol{x}|\omega_i) d\boldsymbol{x} P(\omega_j) \\
&= \sum_{i=1}^{2} \int_{\Omega_i} \sum_{j=1}^{2} C_{ij} p(\boldsymbol{x}|\omega_j) P(\omega_j) d\boldsymbol{x}
\end{aligned}$$

Minimise integrand at all points, choose Ω_1 so

$$\sum_{j=1}^{2} C_{1j} p(\boldsymbol{x}|\omega_j) P(\omega_j) < \sum_{j=1}^{2} C_{2j} p(\boldsymbol{x}|\omega_j) P(\omega_j)$$

In the case that $C_{11} = C_{22} = 0$ we obtain

$$\frac{C_{21} P(\omega_1|\boldsymbol{x})}{C_{12} P(\omega_2|\boldsymbol{x})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Note that decision rule to minimise the Bayes' Risk is the minimum error rule when $C_{12} = C_{21} = 1$ and correct classification has zero cost.

ROC curves

In some problems, such as in medical diagnostics, there is a "target" class that you want to separate from the rest of the population (*i.e.* it is a detection problem). Four types of outcomes can be identified (let class ω_2 be positive, ω_1 be negative)

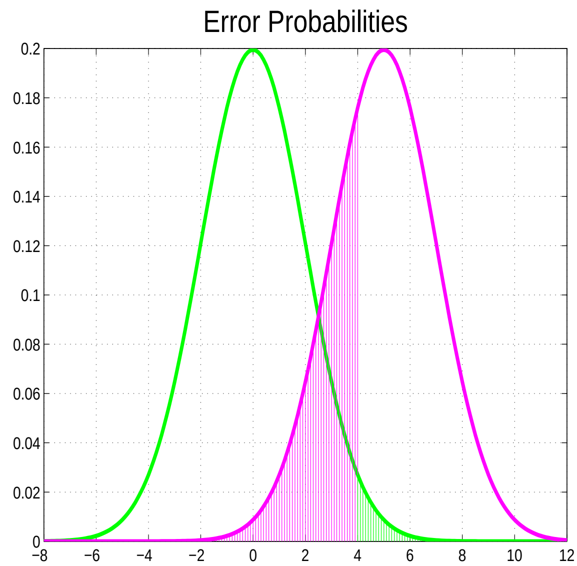
- True Positive (Hit)
- True Negative
- False Positive (False Alarm)
- False Negative

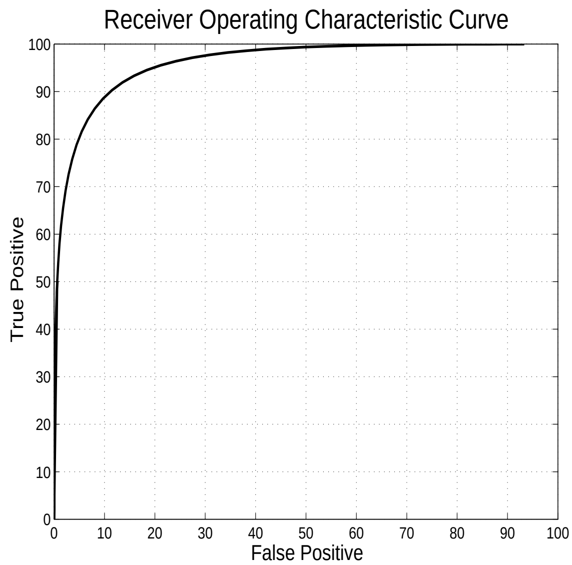
As the decision threshold is changed the ratio of True Positive to False Positive changes. This trade-off is often plotted in a Receiver Operating Characteristic or ROC curve.

The ROC curve is a plot of probability of true positive (hit) against probability of False Positive (false alarm). This allows a designer to see an overview of the characteristics of a system.

ROC curves (Example)

Example 1-d data, equal variances and equal priors: the threshold for minimum error would be $(\mu_1 + \mu_2)/2$.





- Left are the plots of $p(x|\omega_i)$ for classes ω_2 and ω_1.
  - each value of x gives there is a probability for each outcome.
  - for $x = 4$ the probabilities are shown
- Right is the associated ROC curve obtained by varying x (here % rather probability is given on the axis).
  - curves going into the top left corner are good
  - a straight line at 45 degrees is random