Basic probability theory for speech recognition

Objectives

- Understand basic ideas and terminology from probability theory which are relevant to speech recognition
- *Notes: Appendix A, pp 75-92*

The prior probability of a word

- Suppose w is a word
- The probability $P(w)$ is called the **prior probability** (or ***a priori*** **probability**) of the word w
- $P(w)$ is the probability of the word **before** any measurements have been made (hence 'prior')

Random variables

- If every utterance of w resulted in exactly the same signal s speech recognition would be simple
- Unfortunately this is not the case!
- The function, f say, which maps classes to physical measurements, $f(w)=s$, is **not well-defined**, because for fixed w there are lots of possible values of s
- The mathematical notion which was invented to address this problem is the concept of a **random variable**

Random Variables

- By saying that f is a random variable we are acknowledging the fact that $f(w)$ can take on many different values, and we are assuming that these values are determined by a **probability distribution**

Vector valued random variables

- In speech recognition, random variables normally associate classes with **vectors**, rather than with scalars.
- Need to be able to deal with **vector valued random variables**

Probability density functions

- Suppose w is a class, with associated random variable f, and x is a scalar measurement.
- Intuitively, we would like to consider $P(f(w)=x)$
- But $P(f(w)=x)=0$
- Instead we consider $P(a \leq f(w) \leq b)$, **the 'probability that $f(w)$ lies between a and b'** rather than $P(f(w)=x)$, 'the probability that $f(w)$ equals x'.
- To do this we need the notion of a **probability density function** or **PDF**

Probability density functions

- For random variables which take scalar values, a **probability density function** (**PDF**) is just a function p defined on the real line such that

$$p(x) \geq 0 \quad \text{for all real numbers } x$$
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

- If a random variable f is defined by the PDF p then:

$$P(a \leq f(x) \leq b) = \int_a^b p(x)dx$$

Expected Value of a RV

- If a random variable f is governed by a PDF p, then the **expected value** of f, denoted by $E[f]$, is given by

$$E[f] = \int_x xp(x)dx$$

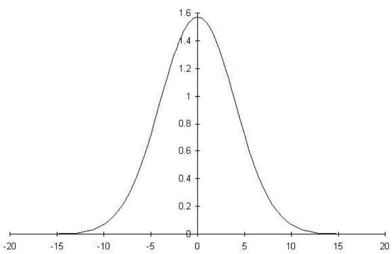
The Gaussian distribution

- A (1 dimensional) **Gaussian** PDF, with **mean** μ and **variance** σ^2 is the PDF p defined by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right]$$

- In this case write $p(x)=N_{(\mu,\sigma 2)}(x)$

Gaussian PDF



Gaussian PDF with mean 0 and variance 4

Multivariate Gaussian PDF

- Consider N scalar valued random variables (one for each dimension) $f_1,\ldots,f_N$, each conforming to a Gaussian distribution with mean μ_n
- If PDFs are mutually independent, their **joint** density is the product of the individual densities:

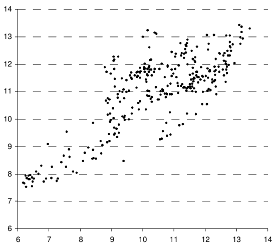
$$p(x_1,\ldots,x_N) = \prod_{n=1}^{N} p(x_n) = \frac{1}{(2\pi)^{N/2}} \exp\left\{ -\frac{1}{2}\sum_{n=1}^{N}(\mu_n - x_n)^2 \right\}$$

Matrix notation

- In matrix notation, writing $x = [x_1,\ldots,x_N]$ and $\mu = [\mu_1,\ldots,\mu_N]$, this can be written as:

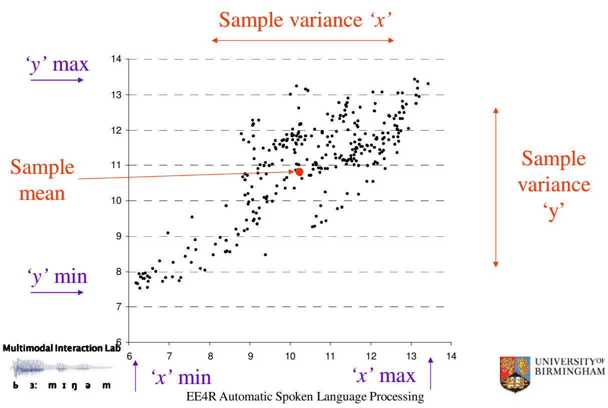
$$p(x) = \frac{1}{(2\pi)^{\frac{N}{2}}|I|}\exp\left[-\frac{1}{2}(x-\mu)^T I(x-\mu)\right]$$

Example from speech processing



Plot of high-frequency energy vs low-frequency energy, for 25 ms speech segments, sampled every 10ms

Basic statistics



Basic statistics

- Denote samples by
  $X = x_1, x_2, \ldots, x_T$,
  where $x_t = (x_t^1, x_t^2, \ldots, x_t^N)$
- The sample mean $\mu(X)$ is given by:

$$\mu(X)^n = \frac{1}{T}\sum_{t=1}^{T} x_t^n$$

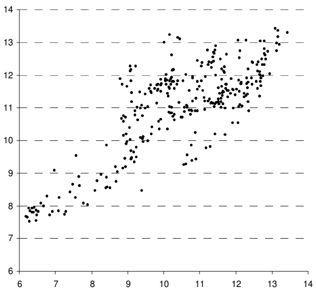
More basic statistics

- The sample variance $\sigma(X)$ is given by:

$$\Sigma(X)^n = \frac{1}{T}\sum_{t=1}^{T}\left(x_t^n - \mu(X)^n\right)^2$$

Covariance

- As the x value increases, the y value also increases
- This is (positive) <u>co</u>-variance
- If y decreases as x increases, the result is negative covariance



Definition of covariance

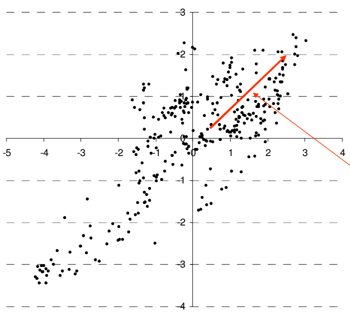
- The covariance between the m^{th} and n^{th} components of the sample data is defined by:

$$\Sigma(X)^{m,n} = \frac{1}{T}\sum_{t=1}^{T}\left(x_t^m - \mu(X)^m\right)\left(x_t^n - \mu(X)^n\right)$$

- In practice it is useful to subtract the mean $\mu(X)$ from each of the data points x_t. The sample mean is then 0 and

$$\Sigma(X)^{m,n} = \frac{1}{T}\sum_{t=1}^{T} x_t^m x_t^n$$

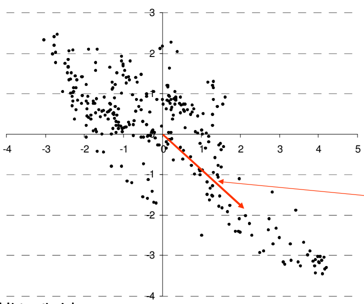
Data with mean subtracted



$$\Sigma(X) = \begin{bmatrix} 2.96 & 1.9 \\ \boxed{1.9} & 1.97 \end{bmatrix}$$

Implies positive covariance

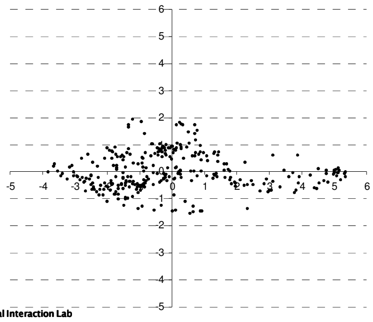
Sample data rotated through 2π



$$\Sigma(X) = \begin{bmatrix} 2.96 & -1.9 \\ \boxed{-1.9} & 1.97 \end{bmatrix}$$

Implies negative covariance

Data with covariance removed



$$\Sigma(X)=\begin{bmatrix}4.51 & 0\\ 0 & 0.48\end{bmatrix}$$

Multivariate Gaussian PDF

- In general:

$$p(x)=\frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\boxed{\Sigma^{-1}}(x-\mu)\right]$$

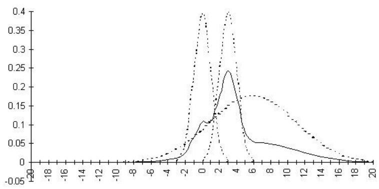
Inverse covariance matrix

- Lesson: Use uncorrelated data if possible

Gaussian mixture PDFs

- An ***M*-component Gaussian mixture** density *p* has the form

$$p(x)=\sum_{m=1}^{M}w_m p_m(x) \qquad 0\le w_m\le 1, \sum_{m=1}^{M}w_m=1 \text{ and } p_m=N_{(\mu_m,\Sigma_m)}$$



Class-conditional PDFs

- Interested in probability $P(a \le f(w) \le b)$, **given the identity of the class *w***
- Write $P(a \le x \le b)$ instead of $P(a \le f(w) \le b)$
- The **conditional probability** that x lies between a and b, **given** that x belongs to class w, is denoted by $\boldsymbol{P(a \le x \le b|w)}$
- $P(a \le x \le b|w)$ is referred to as the **class conditional probability**
- Corresponding PDF is denoted by $p(x|w)$ and is called the **class conditional PDF for the class *w***.

Posterior probabilities

- The probability of the class w **given that the measurement *x* has been observed** is called **posterior probability of the class *w***
- It is denoted by $P(w|x)$

Bayes' Theorem

- **Bayes' Theorem** brings all of these types of probability together
- The form of Bayes' Theorem which we need for pattern recognition is:

Class-conditional density

Prior probability

$$P(w \mid x) = \frac{p(x \mid w) P(w)}{p(x)}$$

Posterior probability

Classification problems

- Suppose we have a finite number of classes, $w_1,\ldots,w_C$ and the goal is to decide which class has given rise to the measurement x
- Since $p(x)$ is independent of w_c, we can ignore it and just maximise the **numerator** in Bayes theorem. I.e:

$$P(w_c \mid x) \propto p(x \mid w_c) P(w_c)$$

Verification problems

- Corresponding **verification** problem is "was the word w spoken?"
- Concerned with the **absolute** value of the probability $P(w_c|x)$
- Need **numerator** and **denominator** in Bayes' Theorem

$$P(w \mid x) = \frac{p(x \mid w) P(w)}{p(x)}$$

Calculation of $p(x)$

- Note that x must be an instantiation of one of the classes, and all of the classes are mutually exclusive.
- Hence a basic rule for calculating probabilities applies and we can write

$$p(x) = \sum_{c=1}^{C} p(x \mid w_c) P(w_c)$$

- So, Bayes' theorem becomes:

$$P(w_c \mid x) = \frac{p(x \mid w_c) P(w_c)}{\sum_{c=1}^{C} p(x \mid w_c) P(w_c)}$$

Parameter Estimation

- Need to estimate the probability density $p(x|w)$ and probability $P(w)$, on the RHS of Bayes' Theorem
- For $p(x|w)$ will use Gaussian mixture PDF, determined by its **parameters**
- Denote the set of parameters by φ
- Once parameters φ are fixed, we write

$$p(x \mid w) = p(x \mid \varphi)$$

- How do we choose the 'best set of parameters' φ ?

Maximum Likelihood (ML) estimation

- Given $x_1,\ldots,x_S$ from class w, assume x_ss are independent and find parameters φ which maximise

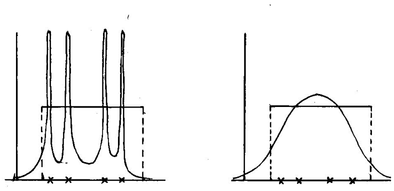
$$L(\varphi) = p(x_1,\ldots,x_s \mid \varphi) = \prod_{s=1}^{S} p(x_s \mid \varphi)$$

- $L(\varphi) = p(x_s|\varphi)$, treated as a function of the parameter set φ, is called the **likelihood of** $\boldsymbol{\varphi}$
- Choosing φ which maximises $L(\varphi)$ is called **Maximum Likelihood (ML) estimation of** φ

Under Training

- A major practical problem in maximum likelihood parameter estimation is **under training**
- Suppose a class w gives rise to measurements uniformly distributed over the interval [0,1].
- Unfortunately we don't know this and try to model the distribution using a Gaussian mixture PDF.
- First, we obtain a training set of S samples $x_1,\ldots,x_S$
- Suppose $S=4$

Under training (continued)

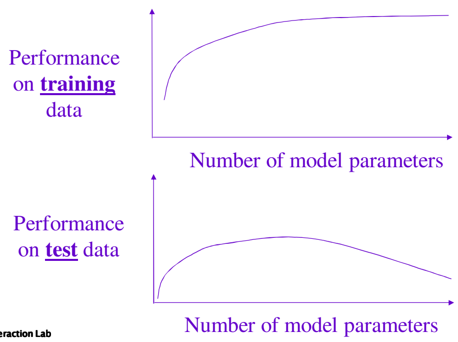


- 4 component PDF gives best fit to training data, but will not generalise to unseen test data
- 1 component PDF performs **worse** on training data, but is a better model!

Under training

- Given a finite training set *X*, and a ML estimate *M* of the parameters of a model, $p(X|M)$ will increase, in general, as the number of parameters in *M* increases
- As number of parameters increases, model begins to characterise detail in the training set which is **not** present in unseen data. The model begins to “remember the training set”
- As number of parameters increases, performance on test data will improve at first, but will then start to degrade as the number of parameters increases and the model focuses on specific detail in the training set

Under training



Experimental method

- Available data is divided into 3 sets:
  - the **training set**, the **evaluation set** and the **test set**
- For each number of parameters, the ML estimate of the parameters is made using the **training** set
- Classification experiments are run on the **evaluation** set, and the number of parameters which gives best performance is chosen for the final system
- This system is evaluated using the **test** set

Summary

- Introduction to basic probability theory
- Random variables, probability functions, and probability density functions
- Gaussian PDFs and Gaussian mixture PDFs
- Posterior, class conditional and prior probabilities
- Maximum likelihood parameter estimations
- Under training