

CFCS1: Assignment 3

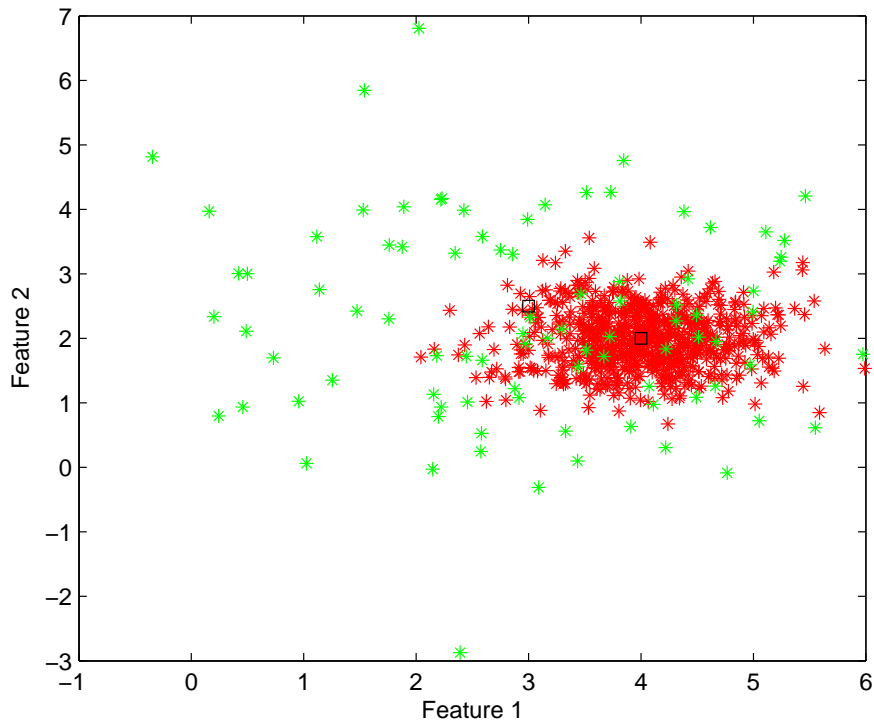
February 22, 2010

Handed out:	26 February 2010
Due date:	8 March 2010

Introduction

This assignment further extends the vowel classification task of Assignments 1 and 2. You will use methods from probability theory to improve classification performance when the samples from different vowels overlap, or have a different shape in the 2-dimensional feature space.

This time, we focus attention on distinguishing between ‘ae’ (labelled 1) and ‘ih’ (labelled 2). 800 samples of these vowels are shown on the following scatter plot, with ‘ae’ coloured red and ‘ih’ coloured green, as before. (The samples have been multiplied roughly by 10 compared to assignment 1).



The means of the two samples (ie. the centres of the clusters) are given by

$$\mathbf{m1} = [6.0; 3.0] \quad \mathbf{m2} = [5.0; 3.5]$$

and these are shown on the plot as black squares.

To load the samples into Matlab, save the file `a3_data.mat` to your working directory. This will create two variables:

- A matrix, `samples`. Each **column**, i , obtained by `samples(:,i)` is the 2-dimensional vector corresponding to the i th sample. (This is the same orientation used in Assignment 2).
- A row vector, `labels`, giving the correct vowel group of each sample.

1 Baseline performance

You should check how well the vowels can be classified using the simple Euclidean distance measure, as used in Assignment 1. This gives a baseline score to improve upon.

To do this, copy `ClassifySamples.m` to your working directory. This function classifies each sample as 1 or 2 depending on which of `m1` or `m2` it is closest to, and returns a count of how many samples were correctly classified. It can be called by typing `ClassifySamples(samples, labels)`.

2 Classification using Bayes' Theorem

Your aim will be to classify each sample according to which vowel is the most likely, given the sample data. In this section you will work through a simpler, artificial example, to give you an understanding of how the method works.

Suppose there are two bags, bag A and bag B, each containing three balls. Bag A contains three red balls and Bag B contains two red balls and one green ball. One of the bags is selected at random and two balls are taken out. Both are red. Just as the two features can be used to give information about which vowel a sample is likely to be, so the colour of the two balls can be used to give information about which bag has been chosen.

1. On paper, calculate the probability that two red balls were drawn:

- (a) given that Bag A was chosen, $p(RR|A)$

- (b) given that Bag B was chosen, $p(RR|B)$

[4]

2. With this answer, use Bayes' theorem to calculate $p(A|RR)$ and $p(B|RR)$, the probability that Bags A and B were chosen, given that the two red balls were drawn. Which bag is most likely to have been chosen?

[4]

Now suppose that there are four bags. One of them is identical to bag A from before; three of them are identical to bag B. Again, one of them is randomly selected.

3. Using a similar method to Question 2 calculate the new value of $p(A|RR)$. Briefly explain why there is a difference in the answer compared to Question 2.

[3]

3 Probabilistic Vowel Classification

The normal distribution is a good model for the distribution of the vowel features. Furthermore, it is discovered that for both the vowels, features 1 and 2 are distributed *independently*. (This was not the case for the vowels in Assignment 2, where the features were closely related).

Cognitive scientists have found that following data about the feature distribution for each vowel:

		Mean (μ)	Variance (σ^2)
ae	feature 1 (x)	6.0	0.3
	feature 2 (y)	3.0	0.2
ih	feature 1 (x)	5.0	2.0
	feature 2 (y)	3.5	2.0

You can see that both features for ‘ih’ have higher variance than that for ‘ae’, explaining why the samples are ‘ih’ are much more spread out on the scatter plot.

Since the two features are independent for these vowel, the joint probability density functions $p(x, y|ae)$ and $p(x, y|ih)$ can be calculated as the product of two single normal distribution density functions $n(x; \mu, \sigma)$, $n(y; \mu, \sigma)$ with the appropriate mean and variance.

4. In Matlab, write a function to calculate the joint probability density functions of the two features from a sample, given the vowel. (Do not use the built-in `normpdf` function). Use your function to compute, for each vowel, the joint probability of

the sample given by `[5.5; 3.7]`. (You should find that the answer is 0.1258 for ‘ae’ and 0.0740 for ‘ih’)

[5]

5. Write a second function to calculate $p(\text{ae}|x, y)$ and $p(\text{ih}|x, y)$ using Bayes theorem, assuming that each vowel is equally common. Then modify `ClassifySamples.m` to classify each sample according to which of these probabilities is the highest (rather than using the distance measure). How many vowels are correctly classified?

[5]

In fact, you can see from the scatter plot that ‘ae’ vowels occur much more often than ‘ih’. This can be taken into account to improve the classifier still further.

6. It is found that ‘ae’ vowels occur with probability 0.75 and ‘ih’ with probability 0.25. Modify your functions used to compute $p(\text{ae}|x, y)$ and $p(\text{ih}|x, y)$. Again classify the samples based on which vowel is more likely, using the new probabilities. How many vowels are correctly classified now?

[4]

To submit

For the assignment, you should submit:

- Written answers to the required questions.
- Printouts of all the Matlab functions you wrote.
- A (brief) transcription of your Matlab session, together with the output that you used to answer the questions.

Please hand in a hard copies of the required material by 4:00pm on the due date to the Informatics Teaching Organisation, Level 4, Appleton Tower. If you have questions regarding the assignment, please contact Yansong Feng at yansong.feng@ed.ac.uk.