Department of Mathematics and Computer Science
University of Southern Denmark, Odense

February 19, 2013
Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 7, Spring 2013

Exercise 1 – Linear discriminants

1. Develop analytically the formulas of a generative algorithm with Gaussian likelihood for a k-way classification problem. In particular, estimate the model parameters.

2. Derive the explicit formula of the decision boundaries in the case of two predictor variables.

3. Implement the analysis in R using the data:

```
Iris <- data.frame(cbind(iris[,c(2,3)], Sp = rep(c("s","c","v"), rep
    (50,3))))
train <- sample(1:150, 75)
table(Iris$Sp[train])
```

   Plot the contour of the Gaussian distribution and linear discriminant

4. Compare your results with those of the `lda` function from the package MASS in R.

   Deepening: read section 4.3.3 of B2 and inspect the outcome of `lda` when run on the full data with all 4 predictors, ie:

```
Iris <- data.frame(cbind(iris, Sp = rep(c("s","c","v"), rep(50,3))))
z <- lda(Sp ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
    ,
         Iris, prior = c(1,1,1)/3, subset = train)
# predict(z, Iris[-train, ])$class

plot(z,dimen=1)
plot(z,type="density",dimen=1)
plot(z,dimen=2)
```

Exercise 2 – Naive Bayes

You decide to make a text classifier. To begin with you attempt to classify documents as either sport or politics. You decide to represent each document as a (row) vector of attributes describing the presence or absence of words.

$$\vec{x} = (\text{goal, football, golf, defence, offence, wicket, office, strategy})$$

Training data from sport documents and from politics documents is represented below using a matrix in which each row represents a (row) vector of the 8 attributes.

$$\mathbf{x}_{\text{politics}} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{x}_{\text{sport}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Using a Naive Bayes classifier, what is the probability that the document $\vec{x} = (1, 0, 0, 1, 1, 1, 1, 0)$ is about politics?