Machine Learning 1
Data Science

Summer Semester 2018

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 11
Classification: Linear and quadratic discriminant analysis

Exercise 1

Let Y be a random variable taking values 0 or 1, dependent on a predictor variable x.

If $Y = 0$ then $X|Y = 0$ is $N(4, 1)$ distributed, and if $Y = 1$ then $X|Y = 1$ is $N(5, 1)$ distributed. We assume that if x is unknown then $P(Y=0) = P(Y=1) = 0.5$

(a) Write down the expression for $\pi_1(x) = P(Y=1|x)$ and simplify as much as possible.

(b) Show that Bayes classifier corresponds to: classify Y equal to one if and only if $P(Y=1|x) > P(Y=0|x)$

(c) Use your answer from part (a) to write $P(Y=1|x) > P(Y=0|x)$ as an inequality in terms of x. Again simplify as much as possible.

(d) Taking the logarithm of this inequality show that the Bayes classifier simplifies to: classify Y equal to one if and only if $x > 4.5$

Exercise 2

Using the same model as in Exercise 1, write an R-Function called `posterior` to compute the posterior distribution of Y given x. You can utilise the function `dnorm(x,mean=,sd=)` to compute the density of a normal distribution. `x` should be an argument to the function `posterior` so your function should use the template

```
posterior<-function(x)
{
?????
}
```

Plot this function using the R function `curve()` using x values from 2 to 7. Check that `posterior(4.5)` gives the expected answer and that this is consistent with the plot.

Now extend adapt function `posterior` to accept the following function arguments with the given default values. After each stage check that your function is giving sensible results by plotting the function.

(a) `pi0` is the prior probability $P(Y=1)$ with default value 0.5

(b) `mu0` and `mu1` are the respective means for group 0 and group 1 with default values 4 and 5.

(c) `sigma` the variance (in both groups) with default value 1.

Exercise 3

In Workshop 10 you fitted logistic regressions to the `Diabetes` data set. Read in this data and define the same training and data set. you will need to load the `MASS` library to have access to the functions `lda` and `qda`.

Fit the following discriminant models, each time plotting the ROC curve and the obtaining the AUC. Use the test data for the ROC and AUC. Display all the ROC curves on one diagram using colour to distinguish the curves.

To fit these models use as a guide the code in James Lab 4.6.3 page 161 and the code hints below.

(a) LDA model using `Age`

(b) LDA model using `BMI`

(c) LDA model using `Age` and `BMI`

(d) QDA model using `Age` and `BMI`

Use the Model with the best AUC to plot BMI against age. Those patients with a predicted posterioir probability can be categorised as High Risk, plot the high risk patients in red.

Code Hints:

```
lda.fit1<-lda(YN~Age,data=train)
p <- predict(lda.fit1, newdata=test)$posterior[,2]
pr <- prediction(p, test$YN)
```