Machine learning II
Master Data Science
Winter Semester 2023/24

Prof. Tim Downie

Workshop 4
Classification: Linear and quadratic discriminant analysis
with solutions

Exercise 1 Bayes Classifier

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

Assume that, 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. The prior probabilities, when x is unknown, are $P(Y{=}0) = P(Y{=}1) = 0.5$

(a) Write down the formula for $\phi_0(x)$, the density of $X|Y{=}0$ and for $\phi_1(x)$, the density of $X|Y{=}1$. Hint: The general formula for a normal distribution can be found on Slide 8 of Lecture 4.

$$\phi_0(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-4)^2\right\}$$

$$\phi_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-5)^2\right\}$$

(b) Write down the expression for the posterior probability $\pi_1(x) = P(Y{=}1|x)$ and simplify as much as possible.

$$\begin{aligned}\pi_1(x) = P(Y{=}1|x) &= \frac{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-5)^2\right\} \cdot 0.5}{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-5)^2\right\} \cdot 0.5 + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-4)^2\right\} \cdot 0.5} \\ &= \frac{\exp\left\{-\frac{1}{2}(x-5)^2\right\}}{\exp\left\{-\frac{1}{2}(x-5)^2\right\} + \exp\left\{-\frac{1}{2}(x-4)^2\right\}}\end{aligned}$$

(c) Check that the Bayes classifier corresponds to:
classify Y equal to one if and only if $P(Y{=}1|x) > P(Y{=}0|x)$. *This follows directly from the definition of a Bayes classifier, "choose the class which maximises the posterior prob" ⇒ Choose $Y = 1$ if $P(Y{=}1|x) > P(Y{=}0|x)$, choose $Y = 0$ if $P(Y{=}0|x) > P(Y{=}1|x)$,*

(d) Use your answer from part (b) to write $P(Y{=}1|x) > P(Y{=}0|x)$ as an inequality in terms of x. Simplify to obtain the inequality

$$\exp\left\{-\frac{1}{2}(x-5)^2\right\} > \exp\left\{-\frac{1}{2}(x-4)^2\right\}$$

$$P(Y{=}1|x) > P(Y{=}0|x)$$

$$\frac{\exp\left\{-\frac{1}{2}(x-5)^2\right\}}{\exp\left\{-\frac{1}{2}(x-5)^2\right\}+\exp\left\{-\frac{1}{2}(x-4)^2\right\}} > \frac{\exp\left\{-\frac{1}{2}(x-4)^2\right\}}{\exp\left\{-\frac{1}{2}(x-5)^2\right\}+\exp\left\{-\frac{1}{2}(x-4)^2\right\}}$$

Multiply both sides by the common denominator, which is positive, to give:

$$\exp\left\{-\frac{1}{2}(x-5)^2\right\} > \exp\left\{-\frac{1}{2}(x-4)^2\right\}$$

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

Taking logs:

$$-\frac{1}{2}(x-5)^2 > -\frac{1}{2}(x-4)^2$$

Multiplying by $-\frac{1}{2}$ means the inequality flips direction:

$$(x-5)^2 < (x-4)^2$$
$$x^2 - 10x + 25 < x^2 - 8x + 16$$
$$9 < 2x$$
$$x > 4.5$$

Exercise 2 Coding the posterior probaility as a function of x

Use your answer from Exercise 1 Part (b) to write an R Function called `posterior` to compute the posterior probability of $P(Y{=}1|x)$. You will start by assuming the same model as in Ex 1, and then generalise it to general π_0, μ_0, μ_1 and σ.

You can use 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 following template

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

Plot the function using the R-function `curve()` for x values from 2 to 7.

```
> curve(posterior,2,7)
```

In Exercise 1 you showed that the most-likely-outcome changes at the point x=4.5.
Use `posterior(4.5)` to find the posterior probability at x=4.5. Why is this result "obvious"?

Now adapt you function `posterior` to accept the following *function arguments* with the given default values.

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

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

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

Check that your function gives sensible results by plotting the function with different argument values. E.g.

```
> curve(posterior(x,pi0=0.95,mu0=11,mu1=10),6,17)
```

Exercise 3 LDA and QDA with the Diabetes Data

In Week 1 you loaded the NHANES data set and used Gibbs sampling to impute the missing values. In ML 1 Worksheet 8 you used a simplified version of this data set which has 3 variables YN (has diabetes or not), Age and BMI. There are 9629 with no missing values. In that Workshop you fitted a logistic regression model to these data. Download the dataset from Moodle.

Also download the template R-File `LDA_Workshop_Ex3.R`. This reads in the data, splits the data into the same training and test data sets as last time. The `MASS` library contains the functions `lda` and `qda`. The template file has the code to fit the LDA model part (a) below, plots the ROC curve and calculates the AUC.

Adapt the code to fit the following four discriminant models, each time plotting the ROC curve and the obtaining the AUC.

Which model gives the best AUC on the test data?

The template code should be enough for you to fit the LDA and QDA models, but further help can be found in Labs 7.4.3 & 7.4.4 in James et. al, pages 177 to 180.

(a) LDA model using `Age`

(b) LDA model using `BMI`

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

(d) Compare the LDA model with the logistic regression model using `Age` and `BMI`

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