Practical 13

Jumping Rivers

Logistic Regression

Question 1

We're going to the model the gender of a patient using their head size. The following code will set up the X_train and y_train objects for you

```
import pandas as pd
import jrpyml
hd = jrpyml.datasets.load_head_size()
X = hd.drop(columns = "gender")
y_train = hd["gender"]
X_train = hd[["head_size"]]
```

a) The following code will show you a boxplot of head sizes per gender. What does this tell you about the relationship between gender and head size?

```
sns.boxplot(y = "head_size", x = "gender", data = hd)
```

b) Write a pipeline to standardise the predictor then perform logistic regression. Use that to fit the model

c) If a patient was to have a head size of 3500, what gender would you predict they were? What is the associated probability?

Questions 2

a) For the model in the previous question, what percentage of predictions did you get right in the training data?

b) Of those the model classified as Male, what percentage were actually Male?

c) The following code will set up and perform 10-fold cross validation on the data. How does the average estimate of the accuracy on the test set compare to the accuracy in part a) ?

```
from sklearn.model_selection import cross_validate
from sklearn.metrics import make_scorer, accuracy_score, precision_score, recall_score
import pandas as pd

acc = make_scorer(accuracy_score)
```

```
def precision(y_true,y_pred):
  return precision_score(y_true,y_pred,pos_label = "Male")

def recall(y_true,y_pred):
  return recall_score(y_true, y_pred, pos_label = "Male")

prec = make_scorer(precision)
rec = make_scorer(recall)
output = cross_validate(model,X_train,y_train,scoring={
  'acc' : acc,
  'prec' : prec,
  'rec' : rec
}, cv = 10, return_train_score=False)
```

d) Can you improve your model by including extra terms in your model?

Discriminant Analysis

Question 3

a) Fit a Linear Discriminant Analysis to the head size data to predict gender, using only head size as a predictor variable

b) If a patient was to have a head size of 3500, what gender would you predict they were? What is the associated probability? How does this compare to the logistic regression from question 1?

c) Fit a LDA to all the variables and obtain some cross validation estimates of accuracy, precision and recall. How do they compare to the full logistic regression model?

d) How does QDA compare?

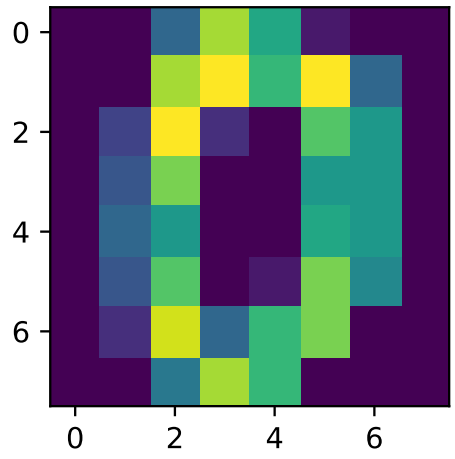
Question 4

The **sklearn** package has some images of hand written digits. It might be interesting to see how well a fairly simple classification model like LDA would do on this. The data can be loaded as below. The plot shows one example image

```
from sklearn.datasets import load_digits

digits = load_digits()
X, y = digits.data, digits.target
```

```
import matplotlib.pyplot as plt
plt.imshow(digits.images[0])
```



a) Break your data into a training and testing split

b) Fit a LDA model to predict the numbers in each image using the training set. You should not need any preprocessing for this

c) How well is the model doing?

KNN

a) Fit a KNN to the digits data, make sure to tune the hyper parameter using accuracy of the classifier as the performance metric.

b) How does this compare to the LDA result?

c) Which digit are we struggling with the most?