Department of Statistics
STATS 330: Statistical Modelling
Assignment 4
Summer School, 2022

Total: 100 marks

Due: 14:00, 8 February 2022

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each question. Include all relevant code and output in the final document.

(iii) Marks may be deducted for poor style. Please keep your code and plots neat.

(iv) These assignments do require you to write R code, but we appreciate this course is not specifically about programming. If you are struggling with the programming aspects of this assignment, please ask for help at our labs or office hours. If you can describe specifically what you want your code to do, then we can point you in the right direction.

(v) Please remember to upload your R Markdown file before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong.

Question 1

Bike sharing systems have become rather common in cities around the world. Using these systems, people are able to rent a bike from a one location and return it to a different place on an as-needed basis. The data set for this question was original provided to a machine learning competition from Capital Bikeshare.[1] We removed some variables to keep the analysis simple. The question is about finding a predictive model for the number of bike rentals in an hour for a given day.

The data set `bike.csv` contains the following variables:

`season`: seasons (1: spring, 2: summer, 3: fall, 4: winter)

`yr`: year (0: 2011, 1: 2012)

`mnth`: month (1 to 12)

`hr`: hour of the day (0 to 23)

`holiday`: whether it is holiday 1 or not 0.

`weekday`: day of the week (0 to 6), i.e. Sunday to Saturday

`weathersit`:
- 1: Clear, Few clouds, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

`windspeed`: Normalized wind speed. The values are divided by 67 (max)

`cnt`: count of total rental bikes including both casual and registered

You may assume the data set is clean for this question.

(a) Load the data set into R, identify and treat all the categorical variables as factors. [3 marks]

(b) Use the code below to do a simple training-test split. Discuss why there is no need to use cross-validation here. State the main reason for avoiding cross-validation. [3 marks]

```
# Split the data
set.seed(123); n = nrow(bike.df); index = sample(1:n)

n.test = round(0.2*n, digits = 0) # 20%-80% split

bike_test.df = bike.df[index[1:n.test], ] # test set
row.names(bike_test.df) = 1:n.test
bike_train.df = bike.df[index[(n.test + 1):n],] # training set
row.names(bike_train.df) = 1:(n-n.test)
```

(c) Fit the full Poisson regression model using the training set. Then use the test set to estimate the mean square prediction error of this model.

[3 marks]

(d) Amongst all sub-models of the full Poisson regression model, use `dredge` in the package `MuMIn` to search for the best predictive model according to BIC. Show the model selection table. Write down the best predictive model in the table according to BIC.

[3 marks]

(e) Use the test set to estimate the Root Mean Squared Logarithmic Error (RMSLE) for each of the first three models in your table for part (d). The RMSLE is calculated as

$$\sqrt{\frac{1}{\texttt{n.test}}\sum_{i=1}^{\texttt{n.test}}(\ln(p_i+1)-\ln(a_i+1))^2}$$

where p_i is the predicted value and a_i is the actual value. Show the three values of RMSLE. Write down the best predictive model out of the three according to RMSLE.

[2 marks]

(f) Use the package `mgcv` to produce a gam plot for your best predictive model in part (e). Based on the gam plot, discuss whether adding a quadratic term is likely to commit over-fitting or under-fitting. Amongst all sub-models of this model with the quadratic term, use `dredge` again to search for the best predictive model according to BIC. Show the model selection table.

[5 marks]

(g) Use the test set to estimate the Root Mean Squared Logarithmic Error (RMSLE) for all the models in your table for part (g). Write down the best predictive model according to your estimated RMSLEs.

[2 marks]

(h) Determine whether the best predictive model in part (h) suffers from under/over-dispersion. If it suffers under/over-dispersion, propose a better predictive model in terms of estimated RMSLE. Justify your answer.

[5 marks]

(i) Use your final model to predict the number of bike rentals between 7am and 8am on a May summer day 2011. Suppose we know it is a working Monday, and the weather is clear with a normalized wind speed of 0.2537. Round your prediction to the closest integer.

[2 marks]

Question 2

The data set `MNIST`[2] is a famous data set of handwritten digits, it is often credited as one of the first data sets to prove the effectiveness of neural networks, it has 70,000 rows and 785 columns. This question is about using it to see how well a logistic regression can predict whether a handwritten digit is a 7 or not.

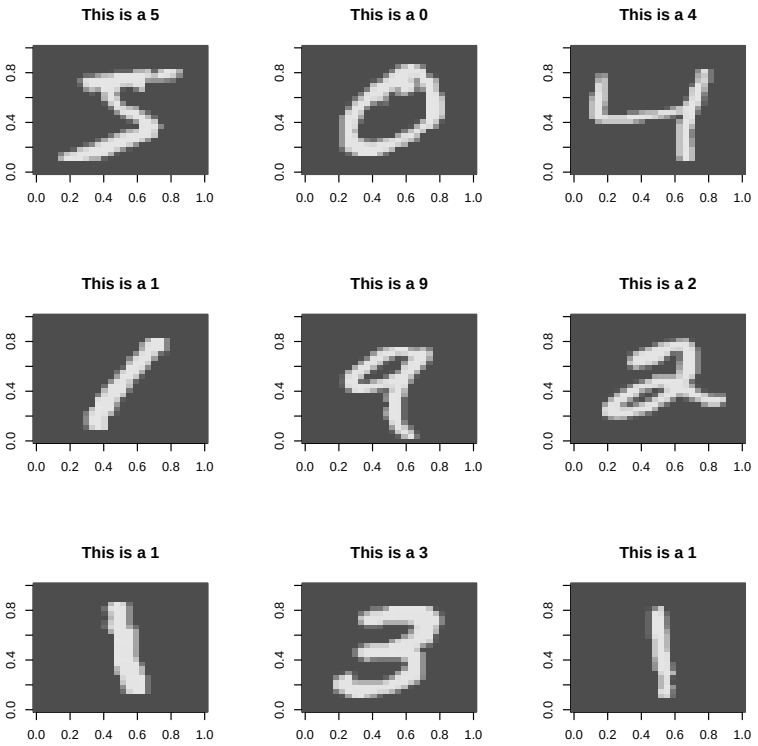

Figure 1: Plot of the first 9 data points.

The data set has been split into training and test sets,

`mnist-train.csv` and `mnist-test.csv`.

Each row in the data sets represents a digital image of a handwritten digit, it is 29 pixels in height and 28 pixels in width, for a total of $29 \times 28 = 784$ pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness/darkness of that pixel, with higher numbers meaning darker. The pixel-value is an integer between 0 and 255, inclusive. The first column is the label for the digit, the rest of the columns contain the pixel-values of the associated image. You may assume the data sets are clean for this question.

(a) Load both the training and test data set into R, then rename the response columns as `Y` and convert the response in both data sets to binary.

$$Y = \begin{cases} 0 & \text{if the digit is NOT 7,} \\ 1 & \text{if the digit is 7.} \end{cases}$$

[3 marks]

(b) Construct the following model. Based on the procedure of predicting 0 if the estimated probability is ≤ 0.5 and predicting 1 otherwise, i.e. cutoff point $c = 0.5$. Produce a confusion matrix of the model for the training data set.

[2 marks]

```
index.predictors = seq(15 * 28 + 1, length.out = 28, by = 1)
predictors.ch = names(mnist_train.df)[index.predictors]
y28.formula = as.formula(
  paste("Y~", paste(predictors.ch, collapse = "+"), sep = ""))
mnist28.glm = glm(y28.formula, family = binomial,
                  data = mnist_train.df)
```

(c) Based on the table in part (b), what is the estimated prediction error?

[2 marks]

(d) Based on the table in part (b), what is the estimated sensitivity?

[2 marks]

(e) Based on the table in part (b), what is the estimated specificity?

[2 marks]

(f) Produce the following plot to illustrate the trade-off between sensitivity and specificity.

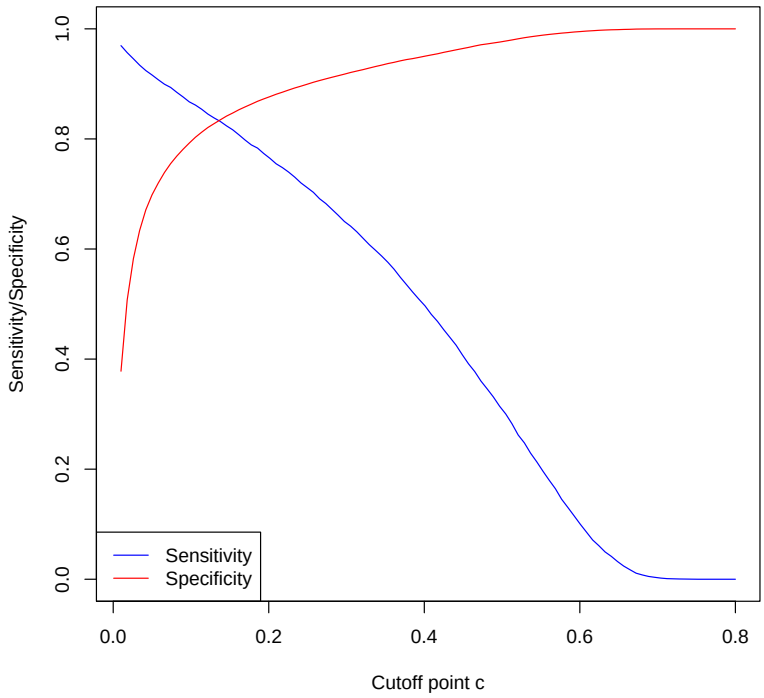

Figure 2: Plot of the first 9 data points.

The values of c used in the plot are given by the following.

[3 marks]

```
n.plot = 100
c.vec = seq(0.01, 0.8, length.out = n.plot)
```

(g) Use `roc` in the package `pROC` to produce a smooth ROC curve plot for `mnist28.glm`.

[2 marks]

(h) What is the area under the ROC curve in part (f)? Explain what does this area represent.

[2 marks]

(i) Find the cutoff point c that maximizes the sum of sensitivity and specificity.

[2 marks]

(j) Find the cutoff point c that maximizes the minimum of sensitivity and specificity. Add this value of c as a black vertical line to the plot in part (f).

[2 marks]

Question 3

Capital Bikeshare, which provided the data set for question one, was interested in the following questions:

I. What seasons of the year are most rides taken on? Is the difference mainly due to the different weather conditions between seasons?

II. How are wind speed and the number of bike rentals related? Is this relationship the same for a holiday and a working day.

III. When do people ride?

Pretend Capital Bikeshare is your client. Your task is to answer those three questions above using the data set `bike.csv`.

(a) Use exploratory data analysis (EDA) to answer question I.

[9 marks]

(b) Use generalized linear models to answer question II. Your analysis should use a plot to illustrate the relationship and how `holiday` affects the relationship.

[15 marks]

(c) Build a model to answer question III. Your analysis may involve data exploration, data wrangling, model building, model checking and interpretation.

[26 marks]

References

[1] Hadi Fanaee-T and Joao Gama. "Event labeling combining ensemble detectors and background knowledge". In: *Progress in Artificial Intelligence* 2.2 (2014), pp. 113–127.

[2] Yann LeCun, Corinna Cortes, and Chris Burges. *MNIST handwritten digit database*. `http://yann.lecun.com/exdb/mnist/`. 2010.