COMM416DAJ – Learning from Data (Professional)

Lecturer: Dr Chico Camargo [f.camargo@exeter.ac.uk]

This assessment will account for **80%** of your final grade for COMM416DAJ.

In this coursework, you will use the tools and techniques you learned throughout this module to train a few machine learning models to predict, classify, and cluster different outcome variables, and discuss the results.

What to submit

You are required to submit your assignment to BART, by **Thursday 8th December 2022, at 11:59am.**
(We might ask you to submit to ELE instead of BART. More updates in the Teams group.)

Please do all your work in a Jupyter notebook. Make a separate cell for every few lines of code, and use separate cells for text. Save your file in the format **COMM416DAJ_STUDENTNUMBER.ipynb**, zip it, and create a file called **COMM416DAJ_STUDENTNUMBER.zip**, replacing STUDENTNUMBER by your student number. For example, if your student number is *34445555*, please submit a file named *COMM416DAJ_34445555.zip*.

This assignment will also use three additional files, named *dataset1.csv*, *dataset2.csv*, and *dataset3.csv*. Do not include them in the COMM416DAJ_STUDENTNUMBER.zip file.

This is an individual assignment. You are welcome to ask for help or assistance on technical issues and general direction of your analysis, and are encouraged to be resourceful when writing your code. In other words: using the internet is allowed, using pieces of code from the practical sessions is allowed, but the overall writing of the code, interpretation of the results, and final output should be your own.

In other words: you are encouraged to search for things like "how to make a scatterplot on python". Just please don't copy code from another student.

Grading

This assignment is worth **100 marks**. The grading will focus on four main points:

1. Is the code syntactically correct? When I open the Jupyter notebook and press "Run", does it run or does it produce an error and stop?

2. Does the code perform the tasks in the assignment?
   - For example: when asked to fit a linear regression, does your code do that?

3. Are the plots clear? Are they readable, and do they convey the message you want to convey?
   - Note that *clear* does not mean *beautiful.* I will only be assessing whether the plots present all the information asked in the assignment in a clear way.

4. Are the questions answered correctly?

Part A – Exploratory data analysis and data visualisation

Economists Serra Boranbay-Akan and Carmine Guerriero have built a dataset describing the locations and operating years of more than 3,000 Cistercian and Franciscan monasteries in 90 European regions between the years 1000 and 1600. This is available in the file *dataset1.csv*, and comes from this article:

> Boranbay, Serra, and Carmine Guerriero. "A novel dataset on a culture of cooperation and inclusive political institutions in 90 European historical regions observed between 1000 and 1600." *Data in brief* 27 (2019): 104731.
> https://www.sciencedirect.com/science/article/pii/S2352340919310868

Here we will use the data produced by Boranbay and Guerriero to train **classification algorithms**.

0. Before running any lines of code, please open a cell and run the command `pip list`.
   This will print out the version of every library you have installed, so I can check in case your code doesn’t run on my computer. After that’s done, carry on with the rest of the assignment.

1. Using the `pandas` library, read the file `dataset1.csv` into a dataframe. Print or display the first five rows of the dataset. **[1 mark]**

2. Make a scatterplot from the dataset, with the `Longitude` column on the `x` axis, the `Latitude` column on the `y` axis, and the colour corresponding to the `Monastery_index` column.
   Within the scatterplot function, set the colour map to `’rainbow’`.
   The result should be a scatterplot showing the position of every monastery, with different colours for monasteries with different values in `Monastery_index`. **[1 mark]**

3. Plot a histogram for the monastery starting year of activity, represented in the `Starting` column. Add the title `“Starting year”` to the plot. **[1 mark]**

4. Make a figure with two histograms, still using the `Starting` column:
   a) One histogram representing the Cistercian monasteries, corresponding to rows in the dataframe where the `Monastery` column is equal to `'Cistercians'`.
   b) And one histogram representing the Franciscan monasteries, corresponding to rows in the dataframe where the `Monastery` column is equal to `'Franciscans'`.
   c) Add a legend indicating which histogram corresponds to which type of monastery.
   d) Add the title `“Starting year”` to the figure.
   [2 marks]

5. Repeat numbers **3** and **4**, but for the `Ending` column. Add the title `“Ending year”` to the plot. **[2 marks]**

6. Count the number of times each country appears in the `Country` column. Use that to make a bar plot, showing one bar for each country, with the bar height representing how many times each country appears in the `Country` column. This corresponds to the number of monasteries found in each country.
   [2 marks]

Part B – Training classifiers

We will now train classifier algorithms to predict whether a monastery is Cistercian or Franciscan, depending on its starting and ending years.

7. Define a variable `X` corresponding to the `Starting` and `Ending` columns of the dataset, and a variable `y` corresponding to the `Monastery_index` column. **[1 mark]**

8. Perform a train-test split, separating `X` and `y` into a training test and a test set, leaving 33% of the data in the test set. **[1 mark]**

9. Classification using a Perceptron:
   a) Fit a perceptron to the training data, and use it to predict `y` values for the test set. **[1 mark]**
   b) Calculate the fraction of data points in the test set where the predicted `y` values and the actual `y` values differ. This fraction should be equal to zero if the prediction is perfect, and equal to 1 if the prediction is wrong for 100% of the test set. Print that fraction. **[2 mark]**
   c) Plot a confusion matrix showing how well the classifier performs on the test set. **[1 mark]**
   d) Treating "Franciscan" as "Positive" and "Cistercian" as "Negative", Print out the precision, recall, accuracy and F1 score of the perceptron. **[2 marks]**

10. Classification using Logistic Regression: do the same as the question above, but using the logistic regression classifier. **[4 marks]**

For open-ended questions like the ones below, write your answers in a new cell of code, either as a commented-out line starting with **#**, or as a *markdown* cell. Indicate what question you're answering, by copying the question as well. For example:

```
Question N. (include the text of the question)
Answer: blah blah blah
```

11. Which classifier (Perceptron or Logistic Regression) performed better at this task? Explain how you've arrived at this answer. **[4 marks]**

12. If you run the code from questions 8, 9, and 10 again, do the scores (precision, recall, accuracy, F1) change? Why (or why not)? Explain where those scores come from, and how do they change (or don't change) if you run the code again. **[4 marks]**

Part C – Linear Regression

A team of Portuguese researchers has created a dataset about wine quality. The dataset includes objective measurements (e.g. pH values), as well as a wine quality measure based on the median of at least three evaluations made by wine experts, who graded the wine quality between 0 (very bad) and 10 (excellent). Part of the data is available in the file *dataset2.csv*, from this article:

> P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, Elsevier, 47(4):547-553. ISSN: 0167-9236.
> http://dx.doi.org/10.1016/j.dss.2009.05.016

13. Read the file `dataset2.csv` into a dataframe. Display the first five rows of the dataset. **[1 mark]**

14. Make four scatterplots, with different variables (different columns) on the `x` and `y` axes, with the variable `quality` on the colour axis. For each scatterplot, choose a different pair of `x` and `y` variables, which cannot include `quality`. **[1 mark]**

15. Print the Pearson correlation between the pairs of variables you have included in the scatterplots. For every pair of variables, print the names of the variables, and the corresponding correlation. **[1 mark]**

16. Linear regression: **[3 marks]**
    a) Choosing the input variable `X` corresponds to any three columns of the dataset, except for `quality`, and the variable `y` corresponds to the `quality` column.
    b) Fit a linear regression between `X` and `y`.
    c) Print out the R^2 score of the linear regression.
    d) Print out the linear regression model's slope coefficients and intercept.

17. Of the three variables you picked to predict `quality`, is it possible to say which variable is the strongest predictor of the variable `quality`, according to your linear regression model? If not, is there any modification to the model that would make it possible to answer that question? Explain how you arrived at that conclusion. **[5 marks]**

18. K-fold cross-validation **[3 marks]**
    a) Using the same `X` and `y` variables defined in the question above, perform a K-fold cross-validation of the linear regression model, with K = 5.
    b) For each fold, calculate the R^2 score.
    c) Print the mean and standard deviation of the five R^2 scores.

19. If, rather than choosing three columns, you had used all columns to predict `quality`, would the average R^2 score be necessarily higher? Why / why not? In which circumstances would it be higher or not? (Feel free to try it in your code! But please provide a justification for your answer) **[4 marks]**

20. Imagine you have two linear regressions, one with $R^2 = 0.80$ and another with $R^2 = 0.90$. Under what circumstances would the first model be preferable over the second one? What if instead you had two classifiers, one with accuracy = 80% and another with accuracy = 90%? **[5 marks]**

Part D – Clustering

The Fingal County Council, in Ireland, has made available a dataset containing the location of trees in Fingal, along with their species and common names. Part of the data is in the file *dataset3.csv*, from here: https://data.smartdublin.ie/dataset/trees

Here we will use the data produced by the Fingal County Council to train **clustering algorithms**.

21. Using the `pandas` library, read the file `dataset3.csv` into a dataframe. Print or display the first five rows of the dataset. **[1 mark]**

22. K-means clustering:
    a) Define a variable `X` corresponding to the `Longitude` and `Latitude` columns of the dataset.
    b) Using the K-means clustering algorithm and the variable `X`, cluster the trees 3 times, using k = 5, 10, 15.
    c) Make a scatterplot showing the results of each clustering, with one colour for each cluster. Suggestion: use a *categorical* colour map such as `tab10` or `tab20`.

    [2 marks]

23. DBSCAN clustering:
    a) Using the same `X` variable as above, using the DBSCAN clustering algorithm, cluster the trees a total of 4 times, setting the `eps` parameter to 0.001, 0.005, 0.01, 0.05.
    b) Make a scatterplot showing the results of each clustering, with one colour for each cluster. Suggestion: use a *categorical* colour map such as `tab10` or `tab20`.

    [2 marks]

24. Using the Silhouette score, compare the 3 runs of K-means and the 4 runs of DBSCAN. Which one of the 7 runs produces the best clustering, according to the Silhouette score? **[2 marks]**

25. Using the Davies-Bouldin score, compare the 3 runs of K-means and the 4 runs of DBSCAN. Which one of the 7 runs produces the best clustering according to the Davies-Bouldin score? **[2 marks]**

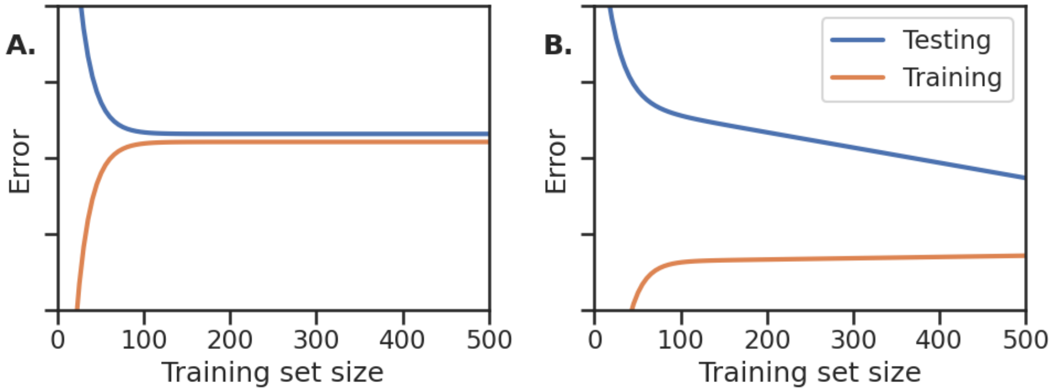
26. Using another clustering algorithm (not K-means, not DBSCAN), and a range of hyperparameter values if appropriate, cluster the trees according to their latitude and longitude, as above. According to the silhouette and Davies-Bouldin scores, does any of your model runs produce better clustering? **[2 marks]**

27. Usually, DBSCAN takes longer than K-means to run, and the time it takes to run is affected by the `eps` parameter. Explain why that is the case. **[4 marks]**

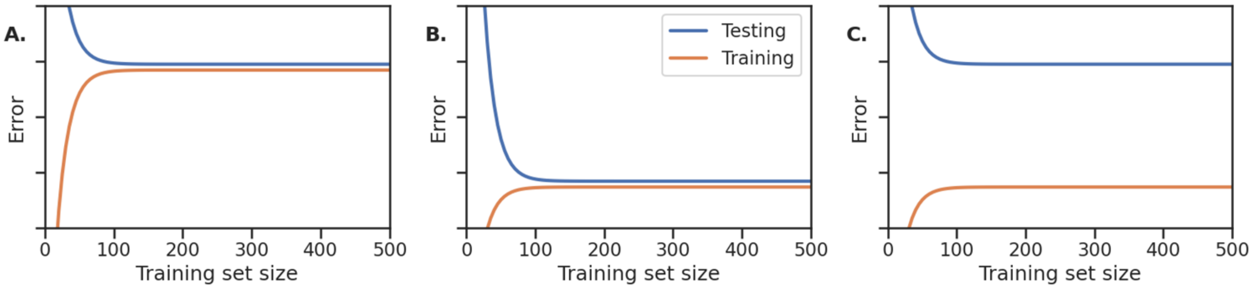
28. Provide an example of one case in which it might be better to use DBSCAN rather than K-means, and an example of one case in which it might be better to use K-means rather than DBSCAN. Explain why, in both cases. **[4 marks]**

Part E – Model selection

For the two questions below, consider that you are training machine learning models after a train-test split. In the figures below, each panel represents a different machine learning model, blue curves represent the error for the Testing set, and orange curves represent the error for the Training set, and the X axis represents the number of datapoints in the training set.



29. Consider the figure above and compare models A and B. In each case, what difference would it make to add more training examples to the training set? Explain your reasoning. **[4 marks]**



30. Consider the figure above and compare models A, B, and C. What is the difference between the three models? How does that relate to the bias-variance trade-off? **[6 marks]**

Part F – Applications of Machine Learning

31. The quality of a classifier can be measured in many ways. Describe three metrics or tools used to measure how good a classifier is, and explain why it might be better to use the three of them, rather than just a single metric. **[5 marks]**

32. Given a classification task and a dataset, sometimes it’s impossible to make a classifier with 100% precision and 100% recall simultaneously. Explain why. **[3 marks]**

33. In the case of the question above, the data scientist might have to choose between having higher precision or higher recall. Provide an example where it’s preferable to get high recall and low precision (and explain why), and another example where it’s preferable to have high precision and low recall (and explain why). **[4 marks]**

Over this module we explained many examples where a machine learning algorithm was trained on a dataset and became reasonably good at a task, but had a fundamental flaw in its training dataset or feature engineering that ultimately made the model inaccurate or inappropriate for use in real life.

34. Give an example of a machine learning algorithm trained for a particular task where it achieves high accuracy in one context, but low accuracy in another context. Explain what could cause that, how to diagnose it, and suggest a way to address it. **[4 marks]**

35. Give an example of a machine learning algorithm that might have low error in its training and testing datasets, but that still would have a fundamental flaw in its application that is not captured by the error metric. Explain why that is the case, and suggest a way to address that. **[3 marks]**