Cx1015 : Online Lab Quiz

- Submit a SINGLE Jupyter Notebook file named **StudentID.ipynb**, where StudentID is your matriculation number.
- Download the **quizData.csv** file posted corresponding to this Lab Quiz. You will need the data for the problems.
- You have **till 11:59 pm, end of day, Thursday, 26 March (today)** to complete the Quiz and submit your solution.

Context

Story. It is quite an important problem in practice to identify "Genuine" bank notes. One may look at the problem from the point of view of *binary classification* in data science or machine learning, where the target is to classify bank notes into two categories – "Genuine" and "Forged". You are given a dataset on bank notes for such a classification problem. Please download the **quizData.csv** file posted corresponding to this Lab Quiz. The problems are all based on this dataset.

Data. The data has been generated by taking a collection of bank notes (both Genuine and Forged) as images, performing wavelet transform on the individual images, and then extracting some statistical features from the wavelet transforms. The resulting features are – **Variance, Skewness, Kurtosis** of the wavelet transformed images, as well as the **Entropy** of the bank note's image. Each feature is "numeric", while the response variable **Banknote** is "categorical" with two levels.

To Do. The goal is to connect the four variables Variance, Skewness, Kurtosis, Entropy to predict the response variable Banknote (Genuine/Forged) using a **Decision Tree Classifier**. In the process, you will also have to do statistical analysis, exploratory data analysis, and other relevant tasks. Plot figures and print outputs as you usually do in a Jupyter Notebook while doing exploratory analysis. If you are supposed to "Comment" something, use the *Markdown* cells in the notebook.

FAQs

This is a Classification Tree problem, and I could not install "graphviz" on my laptop. Will that be a problem?

No, it will not be a problem. Note that you DO NOT have to print the decision tree in any of the problems in this quiz.

What do you mean by statistical description and standard statistical distributions of variables?
What do you mean by "visualize the relationship" between two variables in this context?

Recall/review from the labs the basic statistics and visualizations that are applicable to numeric/categorical variables.

What do you mean by "write a small piece of code" to do something or to print something?

It means that only a few lines of simple code will be sufficient to solve the problem. No need for anything complex.

What do you mean by "show the confusion matrix" in this context?

You may either print the confusion matrix directly (no visualization) or visualize the heatmap of the same. Your choice.

What do you mean by "justify" in certain problems in this quiz?

It means that you will have to write a few lines of justification. No need to write too much. Just the main points.

What do you mean by printing "complete rows from the dataframe" in case of some problems?

Just print the relevant rows of the respective dataframes. You may also want to subset the dataframe and print that.

Will I be awarded part-marks if I can't complete a problem, but do a few parts of it?

Yes. The part marks for each small problem has been clearly indicated. There will be further part marking within.

Problems

Problem 1 : Exploratory Analysis

a) Print the statistical description of the predictor variables in the data and plot standard statistical distributions for each of the predictor variables. The predictors Variance, Skewness, Kurtosis, Entropy are all “numeric”.

b) Comment : Which numeric variable has the most number of outliers? Exactly how many outliers does this variable have, if we consider the points outside the range [Q1 – 1.5 * (Q3 – Q1), Q3 + 1.5 * (Q3 – Q1)] to be the outliers?

c) Print the statistical description and plot standard statistical distributions for the response variable – Banknote.

d) Write a small piece of code to print the exact ratio (“Genuine” : “Forged”) in the response variable Banknote.

e) Visualize the relationship of response variable Banknote with the numeric predictor variables using swarmplots.

(5 + 5 + 3 + 3 + 4) = 20

Problem 2 : Uni-Variate Decision Tree

a) Partition the data randomly into Train and Test sets; 80% for Train and 20% for Test. On the Train set, fit **four uni-variate** Decision Tree models for Banknote against each of the four numeric predictor variables – Variance, Skewness, Kurtosis, Entropy. In each case, do not fit a decision tree more than depth 4 (may overfit otherwise).

b) Predict Banknote using each of the four models on both Train and Test data. Show the Confusion Matrix for each model, both for Train and Test datasets. Print the Classification Accuracy for the tree models, on both datasets.

c) Comment : Which of the four uni-variate Decision Trees is the best in terms of predicting Banknote? Justify.

(4 x 5 + 4 x 4 + 4) = 40

Problem 3 : Multi-Variate Decision Tree

a) Partition the data randomly into Train and Test sets; 80% for Train and 20% for Test (you may use the same partition as before, if you want). On the Train set, fit **a single multi-variate** Decision Tree model (max depth 4) for Banknote against all of the four other numeric predictor variables – Variance, Skewness, Kurtosis, Entropy.

b) Predict Banknote using the tree model you fit on both Train and Test data. Show the Confusion Matrix on both Train and Test datasets. Print the Classification Accuracy for the tree model on both Train and Test datasets.

c) Write a small piece of code to print the exact FPR and FNR values calculated from each of the confusion matrices. You may consider “Genuine” as Positive and “Forged” as Negative in this context and calculate accordingly.

d) Print ONLY the False Positive cases (complete rows from the dataframe) in each of the Train and Test datasets. Print ONLY the False Negative cases (complete rows from the dataframe) in each of the Train and Test datasets.

e) Predict the category of bank note (Genuine/Forged) for the following data points using the tree model you fit.

```
{'Variance': -4.9447, 'Skewness': 3.3005, 'Kurtosis': 1.063, 'Entropy': -1.444}
{'Variance': 0.94225, 'Skewness': 5.8561, 'Kurtosis': 1.8762, 'Entropy': -0.32544}
{'Variance': 2.2429, 'Skewness': -4.1427, 'Kurtosis': 5.2333, 'Entropy': -0.40173}
{'Variance': 0.53936, 'Skewness': 3.8944, 'Kurtosis': -4.8166, 'Entropy': -4.3418}
{'Variance': -2.5724, 'Skewness': -0.95602, 'Kurtosis': 2.7073, 'Entropy': -0.16639}
```

(10 + 5 + 5 + 10 + 10) = 40