COMS4060A/7056A: Assignment #1

Tim Bristow
tim@bristow.za.net

University of the Witwatersrand — September 9, 2021

1 Introduction

This assignment is based on content from Lectures 1-5. You will be required to perform some basic data cleaning and exploration techniques on prescribed datasets. The aim is to explore the dataset and make observations. There is no strict requirement on the format of your submission, but any answers should be reasoned, discussed, and relevant data or results should be provided to substantiate this. You might like to submit either a PDF or a Jupyter notebook, for example. You can use any programming language or tool you would like, however.

1.1 Problem definition

As mentioned in the lectures, data exploration and visualisation is as much an art as it is a science. There is back and forth between the different sections, and the flow need not be linear. There are some points highlighted below that you will need to address for this hand-in, but the order is not important. The most important aspect of this assignment is that you are critical of the data, question your findings, and investigate your data in-depth. Ensure that you understand what the variables represent in your dataset and their datatypes. If there is specific domain knowledge that you find, highlight it and what the consequences are (this isn't required for this assignment though).

Note, this assignment mostly just involves lots of plots - it might appear long, but once you've got the hang of how to produce plots it shouldn't take too long. Make use of code from the Jupyter notebooks that have been provided throughout the lectures.

1.1.1 Data cleaning and outliers

There is a popular dataset (it is an older dataset but it checks out) that contains information on glass identification . There are 214 glass samples split amongst seven class categories and nine features, including the refractive index and the content in percent of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe.

1. Using visualisations, explore the feature variables to understand their distributions as well as the relationships between predictors. Here, include histograms, bar charts, correlation heatmaps, etc. [6]
2. Can you find any outliers? Are any of the distributions of the features skewed? [3]
3. What types of transformations of one (or more) of these features might improve the classification model? [1]

Total Marks: 10

1.1.2 Data cleaning and missing values

There is another popular dataset containing soybean data. The data were initially collected to predict disease in 683 soybeans. The 35 features are mostly categorical and describe the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The target variable has 19 classes.

1. Produce visualisations showing the frequency distributions for the categorical features. Are any of the distributions redundant? [5]
2. Roughly 18% of the data are missing. Are there particular features that are more likely to be missing? Does it appear to be related to the classes? [2]
3. Develop and implement a strategy for handling missing data, either by eliminating predictors or imputation. [5]

Total Marks: 12

1.1.3 Feature Selection and Engineering

IBM has a sample dataset that contains data on customer turn for a telecoms provider. The problem definition from IBM is "Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs."

The dataset is available from Kaggle: telco churn. (If you do not want to create a Kaggle account to get the data I will upload it to Moodle too). IBM subsequently enriched the dataset, and they include some additional commentary and insights here. You do not need to use that dataset, it is purely if you're interested and want more background.

Each row in the dataset represents a customer, and each column contains customer's attributes described below:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

The dataset is pretty clean, and there are only missing values for 11 values of TotalCharges. You can drop these rows.

With this dataset, you need to do the following:

1. Convert categorical features to dummy variables (Yes/No counts as categorical). [1]
2. Plot the different features, including the distribution of the target variable. [5]
3. How are tenure and contract duration related? Show this with a visualisation. [2]
4. Look at the correlation between Churn and the other variables - are there strong negative or positive correlations? (Use f-regression or mutual-info-classif, for example). [5]
5. Produce plots to look at churn vs tenure, contract, age, monthly and total charges. [5]
6. Use logistic regression (l1 norm) and a random forest to get a list of the most important variables. How different are they from each other, and how do these relate to the variables from the correlations above? [10]

Total Marks: 28

1.1.4 Dimensionality reduction: PCA

Use the penguin dataset for this section. For this piece do not worry about cleaning the data or looking for missing values.

1. Perform PCA with 2 and then 4 components. Show the explained variance for the different PCs.
2. Then, for the PCA with 4 components, make a scatterplot for the first two principle components for a) the raw data and b) standardised data. What do you notice about these different plots?

Total Marks: 10

1.2 Submission

Work in groups of up to four people. Submit your work to Moodle as a PDF or Jupyter notebook.

Deadline: 17 September 2021 (please contact me directly regarding extensions)