COSC 2673 Machine Learning

Lab 02

Objective

- Continue to familiarise with Python
- Load dataset and examine the dataset
- Learn to compute basic statistics to understand the dataset more
- Plot the datasets to visually investigate the dataset

Introduction

In this lab, we get some initial experience with using some of the main python tools for this course, including Numpy, Matplotlib and Pandas. We also load some datasets, compute some basic statistics on them and plot them.

Datasets

We examine two regression based datasets in this lab. The first one is to do with house prices, some factors associated with the prices and trying to predict house prices. The second dataset is predicting the amount of share bikes hired every day in Washington D.C., USA, based on time of the year, day of the week and weather factors. These datasets are available in housing.data.csv and bikeShareDay.csv. Next we examine how to load these into Python and Jupyter notebooks. Note you can use other approaches and IDEs if you choose to, but the lab and help are based on Jupyter notebooks.

Loading Dataset

First, ensure the two data files are located within the Jupyter workspace. We will first analyse the Diabetes dataset, then you'll repeat the process to analyse the bike hire dataset.

First in need to import a few packages. Type these into Jupyter notebooks then press run.

```
$ import pandas as pd
$ import matplotlib.pyplot as plt
$ import numpy as np
```

Pandas is a great Python package for loading data. We will use Matplotlib to visualise some of the distributions. Numpy is numeric library that has many useful matrix and mathematical functionality.

Next, we use pandas to load out dataset:

```
$ bostonHouseFrame= pd.read_csv("housing.data.csv", delimiter="\s+")
```

Assuming the dataset is in ‘‘housing.data.csv’. Replace this with the relative or absolute path to your files. We strongly encourage you to look up the documentation of the functions we use in the lab, in this case examine https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html .

The ‘read_csv()’ command loads the input file, which is a csv formatted file, delimited by tabs, into a Pandas dataframe, which can be thought of as a table, which can also store the column names as well as the data. Examine what has been loaded into the dataframe ‘bostonHouseFrame:

```
$ print(bostonHouseFrame)
```

What is the output?

Data frames are a very useful tool, and we strongly suggest to familiarise yourselves with it. Here is some introductory material for it: https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python .

Now we have loaded the data into a data frame and printed it out, next we will compute some very basic statistics.

Examining the Dataset

We compute some basic statistics to examine the values within the dataset. Looking at the maximum, minimum, average and median values for each column (which equates to a feature or attribute) is often an useful way to obtain a quick summary of the data and how it is distributed (which in later labs we will examine if we need to do some pre-processing, and even if not, provides us with some information about potentially which machine learning approach might work well).

First we compute the maximum. There are multiple ways of doing it, but we use Pandas built-in max function:

```
$ pd.DataFrame.max(bostonHouseFrame)
```

Anything in particularly striking?

Compute the minimum now:

```
$ pd.DataFrame.min(bostonHouseFrame)
```

Contrast the min and max values for BMI.

Compute the average (using mean() function):

```
$ pd.DataFrame.mean(bostonHouseFrame)
```

Now compute the median:

```
$ pd.DataFrame.median(bostonHouseFrame)
```

What do you think all these statistics are useful for?

Compute Basic Statistics

In addition to computing the min, max, average and median, we can also examine the distribution of the data. This can be important, to identify obvious skews, to obtain a better understanding of what is going on, to see if the data satisfies the assumptions required for machine learning techniques you'll learn later.

First, we look at the histogram of each of the attribute to examine their distribution:

```
$ plt.figure()
$ bostonHouseFrame.hist()
$ plt.show()
```

Recall a histogram plots the (binned) values that each attribute can take in the x-axis, then the total frequency for each bin in the y-axis. Consider the histogram plots for age and sex. Do they make sense?

Box-plots are another good visualisation to examine the data. They essentially plot the min, max, 25-50-75 quartiles, and can quickly provide a visual summary of the distribution of each feature.

```
$ bostonHouseFrame.plot(kind='box', subplots=True, layout=(4,4), sharex=False,
sharey=False)
$ plt.show()
```

Finally, it is typically useful to see how correlated a pair of features/attributes are. This can tell us which ones are possibly redundant (if two features are highly correlated, i.e., when one increases the other increases (or decreases if anti-correlated) very similarly, so one can be used to predict the other. Type in the code below and run it to get a correlation plot.

```
$ correlations = bostonHouseFrame.corr()
$ fig = plt.figure()
$ ax = fig.add_subplot(111)
$ cax = ax.matshow(correlations, vmin=-1, vmax=1, cmap=plt.cm.PuBu)
$ fig.colorbar(cax)
$ ticks = np.arange(0,14,1)
$ ax.set_xticks(ticks)
$ ax.set_yticks(ticks)
$ ax.set_xticklabels(bostonHouseFrame.columns)
$ ax.set_yticklabels(bostonHouseFrame.columns)
$ plt.show()
```

Which features do you think are correlated?

Exercise: Analyse the Bike Share Data

Now you seen how to do this task for the Diabetes dataset, repeat the same process for the Bike Share data. Going through the same process, please answer the following questions and discuss this with your lab demonstrator (please do attempt this and don't just read the solutions later, as although as a teaching team we don't really mind, but you don't really benefit).

- What is the range of some of the attributes?
- Which of the features have a very different average to the others?
- Which feature is skewed (hint examine the histogram)?
- Which features are highly correlated?

Relate above questions back to the domain of the dataset (bike sharing) and see if you can come up with explanations for the observed data.

Roadmap for next lab

In this lab, we had studied how to load some datasets, examine them and perform basic analysis on them. In the following lab, we will learn how to do regression and examine and evaluate the output.