Machine Learning – Lab

**Unsupervised Learning**

**A. Overview**

In this lab you will apply ***unsupervised learning techniques*** to a dataset consisting of city bus locations.

We have provided two files that contain data reported from buses that operate along the MBTA 1 and 43 bus routes. The data are provided in tab-delimited files. Each row represents data reported from a single bus at one point in time, so there will be many rows that relate to a single trip of the same bus.

**B. Data Cleaning, Analysis, and Visualization**

The file to be used for this section is a tab-delimited file “bus\_data.tab” which contains data reported from buses that operate along the MBTA 1 bus routes, over a 24 hour period. Each row represents data reported from a single bus at one point in time, so there will be many rows that relate to a single trip between Dudley Station and Harvard Square (the two end-points of the bus route.) Each row also includes data about the geographic location of a bus, including its longitude, latitude, and heading, which is the angle the bus is pointing.

1. Import and review data

Load the tab delimited bus data file (‘bus\_data.tab’) and inspect it using the data table widget.

1. Are there any attributes in the dataset that are not useful for analysis or for making predictions? Why?

The Local would be useful but it can't be coerced into a datetime in Orange

Useless: Local, utc, secsSinceReport, speedKmHr, epoch.

They do not seem to be necessary information

1. Do any of the variables appear to be treated as discrete even though they actually represent continuous values?

Anything that is temporal and heading

Hint: You can either view the values in a “Data Table” widget, or use the icon next to each attribute in the “Select Columns” widget to view which attributes Orange thinks are discrete or continuous.

1. Do any of the variables seem to represent the same data?

Epoch, Time, lasttime

1. What do you think the dirTag attribute represents?

and id for the direction

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

1. Inspect the distributions of attributes and selecting data
2. Now inspect the distribution of route numbers. Are there any bus routes that should not be included in our 1 bus analysis? If so, remove those using the select data widget.

66 and 43 has significantly less data

1. Is there anything else we should filter that is not a complete attribute?

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

1. Selecting attributes to analyze
2. Which attributes should be excluded from analysis, and where do you need to put them in the “Select Columns” widget? Are any of them perfectly correlated? Should we include all of those?

Do not exclude: epoch, heading, predictable

1. If you want to ‘label’ points in your analysis with the direction the buses are heading, which attribute do you need to put in the “Target Variable” position?

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

1. Select Data

Use the Select Rows widget to filter out the rows that have irrelevant or undefined values for dirTag, and those that have any routeTag other than 1.

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

1. View the route
2. Choose features that will help you plot the geographic locations of the route on a graph. Save the graph you produce using the “Save Graph” button.
3. How closely does this mesh with the actual bus route? Can you guess what is happening when there are any deviations from the actual route?

the majority of the points match the actual route

deviations may have occurred due to road closings/ construction

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

1. Bus time and frequency

We want to understand when the buses run and their frequency throughout the 24 hour period.

1. Plot out the frequency of the bus observations by two hour increments. Save the graph you produce using the “Save Graph” button.
2. What are the peak periods when the buses run most frequently?

Hint: You can set a widget to view the number of rows that appear by time interval, and then you can make the number of different time intervals 12 instead of the default value, representing 2 hour windows.

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**

**C. PCA**

Perform PCA analysis on this data using the PCA widget.

1. How many principal components account for almost all (about equal to 100%) of the variance? Inspect the components using a Data Table widget. Comment on the features that form the primary “ingredients” in each of these principal components.

6 principal components

Primary ingredients: predictable, routeTag, longitude

1. In one or two words discuss what the first, second and third principal components represent. Discuss why these features dominate the variance.

Primary ingredients: predictable, routeTag, longitude

1. Plot the data as projected onto two principal components using a Scatter Plot widget. Find a scatter plot that well separates the two direction tags. Save the plot with the “Save Graph” button. You should see that most of the points cluster into two mostly rectangular blocks. How do you interpret these blocks? How do you interpret the points that lie outside of these blocks? Hint: Remember what you said these principal components represent in the previous question.

Most of the data is explained by PC1 and PC2 but those that lie outside of the 2 rectangles probably have an unpredictable value, route Tag that is not 1 and longitude values that doesn't lie along bus1

**PAUSE HERE, BRING FORWARD ANY QUESTIONS TO THE CLASS**