ECS607/766 - Data Mining - 2015/16

Lab 1: Data Exploration & Visualization
01/10/2015

Introduction

The outcome from the lab is to be familiar with Weka and basic exploratory data analysis. **This session is a warm-up that will not count toward your final grade.**

1. Getting Started

1. Download the dataset for this week (qmplus: LondonCars.csv)
   (This is data contains all the cars for sale on Autotrader.com. It was collected in Summer 2014 by an MSc student who worked on a data mining project)
2. Invoke Open Office or Excel Spreadsheet
3. Invoke Weka.
   -Windows: Computer → teaching → weka → weka.jar
   -Linux: search → Weka

2. Explore the Dataset – Open Office

1. Load the dataset LondonCars.csv in a Spreadsheet (Open Office or Excel)
2. Basic exploratory questions:
   a. How many instances does the dataset have?
   b. How many attributes? What are they?
   c. What are the possible values for Body Style & External Color?
   d. What is the minimum, maximum, average and median price?
   e. Why might the median price be different than the average price?
   f. What is the most common year of car?
   g. What is the ratio of 2-door to 4-door cars? Hint: Try countif function.
   h. What is the average price of a Honda car versus a Mercedes-Benz car? Hint: Try averageif function.

3. Explore the Dataset - Weka

In this part we will explore the same data with Weka data mining software. Some things will be easier to do here than with a spreadsheet program. This type of exploratory analysis is typical when you are faced with a new dataset that you need to understand before trying to solve a specific problem.

1. Start Weka Explorer
2. Open the dataset: Open → Select csv → LondonCars.csv. You can click edit to

see the raw data in a spreadsheet style.
    a. Again, find out how many instances, and attributes the dataset has?
    b. By clicking on an attribute in the "Attributes" panel, you can see its type, possible values (discrete), or statistics (continuous) in the "Selected Attribute" panel.
    c. Which attributes are continuous (numeric) or discrete (nominal)?
    d. What are the possible body styles? Which is the most and least common body style?
    e. What are the most and least popular external body colors?
    f. What is the minimum, maximum and average values for mileage and price?
    g. Looking at the histogram for those attributes, do they look Gaussian (bell curve) distributed? Are there outliers?
3. If you use the Class pulldown in the Selected attribute panel. You can color the histogram of one attribute according to another attribute.
4. Select Body style attribute and Body style class. Note the color corresponding to each body type.
    a. Select Make attribute, keeping body style coloring.
    b. Which car company only makes SUV style cars?
    c. Which car company makes the most Coupe style cars?
5. The #doors attribute has been interpreted as numeric. How many unique doors are there? To use it as a class for coloring, we should convert it to nominal.
    a. Click Filter → Unsupervised → NumericToNominal. Click the box next to the filter and **choose attribute index = 11** (index of doors) and press apply. Be careful not to apply to all attributes, or they will all become nominal.
    b. Click the doors attribute and observe the coloring of each door configuration.
6. Now select body style attribute and color it by doors class.
    a. Which car types always come with 2 doors? Which always come with 4 doors?
    b. Which car types come in a mix of door configurations?
    c. For each class attribute, you can click Visualize All to see each attribute histogram in those colors. See what other patterns you can find.

4. Finding Correlations

1. Switch to the visualization view in Weka Explorer. This can plot any pair of attributes against each other. Click on any panel to bring up a plot. On these plots every instance (record or row) in the database is shown as a point. If you click on any point it will bring up a window showing you the details of that instance (car).
2. Select X: Mileage, Y: Price. What do you observe about the relation between the two?
    a. Is there any correlation? Is it linear?
    b. You can simultaneously color the plot by any other variable.
    c. What do you observe when coloring by price?

    d. When coloring by Engine size, what is the impact of 8, 6 or 4 cylinder engines?
3. Plot X: External Color, against Y: Price
    a. Which color looks like they get the highest valuation overall?
    b. Which color gets the lowest valuation overall? Use the Jitter slider to slightly spread out all the points that are on top of each other.
4. Looking at the least valuable color:
    a. How many cars with that color are shown?
    b. Click on a car (data point) of that category, what make is it?
    c. Is it safe to conclude that cars with that color are generally very likely to be cheap?

5. Making Predictions

Lets try to build a car-price predictor. This would be useful for a used car business to know how much to offer to pay for a used car, and how much to sell it for.

1. Switch back to the Preprocess view. Remove all attributes besides the numeric ones: Year, Mileage, Price. (Click the nominal attributes and then press Remove). In later exercises we will get back to predictions using nominal inputs.
2. Switch to classify view.
    a. Under classifier. Choose Classifiers → Functions → Linear Regression.
    b. Make sure Price is selected as the target variable to predict in the drop-down box, and press start.
3. Observe the results in the output panel
    a. The regression model discovers the values of A, B, and C in an equation of the form Price = A*Year+B*Mileage+C. It can then use this equation to predict the price of a new car.
    b. What is the value of every Year of car age?
    c. How much value does every mile of driving loose?
    d. The mean absolute error is the average difference between the predicted price of each car and the true price. What is it in this case?
    e. Do you think this an acceptable level of prediction accuracy for a used car business? What could we do to improve the accuracy?