Benjamin Arancibia
2/2/2015
IS-643

Assignment 1: Introduction to Data Mining

1) A typical dataset has several examples (instances); each instance is comprised of a number of values per variable (attributes). Based on this there are two different types of data, supervised learning and unsupervised learning, which are treated in different ways.

Supervised learning is a specifically designated attribute and the aim is to use data to predict the value of that attribute for instances that have not yet been seen. If the designated attribute is categorical with distinct values such as: good, very good, bad, car, and person, the task is called classification. If the data used in supervised learning is numerical such as, expected sales price of a house, the task is called regression.

Unsupervised learning is data that does not have a specifically designated attribute and the hope is to extract the most information from the data available. Unsupervised learning tends to use probabilities and find associations between different variables.

2) A classification problem that is of interest to me is finding apartments in neighborhoods that are likely to rise in value, fall in value, or have unchanged value. The use case of this finding is finding the best apartment for the best price, before the rent increases. At a later point in life this same classification could be used to buy a house.

The necessary data for this classification problem is rent amounts, neighborhood ratings, and transportation data. Rent amounts could be scraped from different apartment hunting sites, but they might differ from website to website and neighborhood ratings is not defined. Based on the end result user's values, the same neighborhood could have different ratings based on the variables values. For a one-off example this will not be an issue. Data types in this example would be Ordinal (apartment size), integer (rent amount), and nominal (neighborhood ratings) categories.

3) An estimation problem that is of interest to me is predicting how well a CPU will perform on a task. The input of the data into the model will result in a numeric performance number.

The necessary data for this estimation problem is the processor's cycle time, minimum amount of main memory used, maximum amount of main memory used, cache, and min/max of CPU channels. Hopefully, this problem could use linear regression to solve the issue.

4)
(a) This scenario is a regression problem and we are more interested in inference. We are looking at ways that Y is affected by the X1…Xn change.
(b) This is a classification problem and we are interested in prediction.
(c) This a regression problem are we are interested in a prediction and inference. We are looking at ways that X1…Xn affects Y and predicted the change.

5) An interesting problem to me is predicting how certain events throughout the world effects prices of goods and commodities. This would require a learning approach and could fall under pattern mining within the umbrella of data mining. For example, how does a terrorist attack in Nigeria combined with an increase of onion prices in Indonesia effect oil prices in Argentina? Finding patterns associated with different events would be an interesting.

6) I expect to use Python and Microsoft Azure Learning. I enjoy Python and like using it with the various libraries that work well with data science concepts. Examples are SciKit including Pandas, numpy, seaborn, etc. I would also expect to use Microsoft Azure Machine learning because I am really interested in doing ML in a more distributed format.

7) Done!