Machine Learning and Data Analysis

Course Description:

Humans are able to learn from experience, and get better at their tasks with more experience, and more data. An important question is whether we can automate this learning process? This is precisely what the science of machine learning aims to do. This science of machine learning is also what drives the applications of most major technology companies, from search engines, to social networks, to being recommended items that we might like. This course will provide a brief introduction to the topic of machine learning and data analysis, with some key methods and principles underlying these methods.

Goals:

1. To gain a solid background in the basics of data analysis and machine learning.
2. To understand how to analyze a given dataset, and predict future outcomes using machine learning techniques.
3. To work on a research project and write a report.

Prerequisites: Proficiency in high-school level calculus, and linear algebra.

Individual Evaluations: All the students will be evaluated on the basis of the following. Class participation will carry 30% weightage. This includes asking questions in the class, answering the questions asked by the Professor, and overall ability to follow the lectures The final project will carry 70% weightage. All students will be required to produce a report which will be evaluated on the basis of the overall quality of work done, demonstration of the understanding of the research topic, and, the quality of the write up in English language.

Course materials:

There is no required textbook, and the lectures slides will be self-contained. The following textbook might be useful for supplementary reading: Introduction to Data Mining, by P. Tan, M. Steinbach, V. Kumar, Addison Wesley, 2006. Textbook Website: http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

List of topics:

Topics covered will include exploratory data analysis, clustering, and prediction using decision trees.

Research project:

All students will be required to work on a research project from one of the following three topics: Exploratory Data Analysis and Data Visualization, Clustering and Grouping of Data, Prediction using Decision Trees.

Course Schedule:

Following is the tentative course schedule.

The first five weeks consist of the following **lectures**.

Lecture 1: Types of Data, and Analysis of Data

Lecture 2: Exploratory Data Analysis (EDA)

Lecture 3: Clustering I

Lecture 4: Clustering II

Lecture 5: Decision Trees

The next five weeks will consist of the **research project.**

The duration of each project will be 5 weeks. First week will be dedicated simply to understanding the problem and the background material. Weeks 2 to 4 will be used to carry out the required research. This will involve consulting various books, internet sources, and, published articles. The last week will be used to write the final report and polish it as much as possible.

The student can choose from one of the following three topics.

I. Exploratory Data Analysis and Data Visualization

Here we aim to "understand" a given dataset by asking various scientific questions about the dataset, and answering it via elementary data analysis as well as data visualization techniques. It trains to think like a scientist, specifically, a data scientist.

As a preliminary task, you will download the Boston housing dataset from https://www.kaggle.com/c/boston-housing/data

You will then perform an exploratory data analysis on this dataset, answering the following questions;

1. How many rows and columns are there in this data set? What do the rows and columns represent?
2. Make some pairwise scatterplots of the attributes (columns) in this data set. Describe your findings.
3. From the scatterplots, can you identify any attribute that is associated with per-capita crime rate?
4. Do any of the suburbs of Boston appear to have particularly high crime rates? Do any have particularly high pupil-teacher ratios?
5. How many suburbs in this data set bound the Charles river?
6. What is the median pupil-teacher ratio among the towns in this data set?
7. In this dataset, which suburb of Boston has the lowest median value of owner-occupied homes? What are the values of the other attributes for this suburb, and how do those values compare to the overall range for those attributes?

8. In this dataset, how many of the suburbs average more than eight rooms per dwelling? What can you say about these suburbs?

Finally, you will then download some additional UCI datasets from https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=less10&numIns=&type=&sort=nameUp&view=table

and ask and answer questions similar to above.

II. Clustering and Grouping of Data:

Here we aim to understand a dataset by grouping together the data samples into a small set of groups. For instance, in a classroom with 100 students, suppose we want to group these students into three groups. How should we do so?

In this project, you will pick three UCI clustering datasets from https://archive.ics.uci.edu/ml/datasets.html?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table

and obtain interesting clusterings or groupings of the corresponding datasets. You will be contrasting two clustering techniques: K-Means, with different initialization heuristics, and hierarchical clustering.

III. Prediction using Decision Trees

One of the most important machine learning tasks is to predict some outcome. For instance, in the stock market, we might be interested in predicting the future stock price of some company.

Or for an e-commerce company might want to predict if some user will like to buy a particular product or not.

In this project you will pick three UCI classification datasets from https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table

and test predicting the corresponding outcomes detailed in the datasets, using the machine learning technique of decision trees. You will be comparing different splitting approaches to learn decision trees.