Introduction to Machine Learning: Prof. Danna Gurari

Problem Set 1

Question1

Supervised versus Unsupervised Learning [7 points]: In your own words, define each learning approach, describe the difference between them, and name two commercial applications learned using each approach (i.e., 4 applications in total).

Answer –
Supervised Learning – The approach in which each observation of the predictor measurement (say x1, x2,…xi) there are an associated response measurements (say y1,y2,….yi) respectively are available and we wish to fit a model that relates to a response to the predictors, with aim of accurately predicting the response for future observations (Prediction problem) or better understanding of the relationship between the response and the predictors (Inference problem) is called as Supervised Learning approach.

Unsupervised Learning – Unsupervised learning is somewhat more challenging than Supervised Learning, as for every observation (say x1, x2,…xi) there is no associated response (y1, y2, …..yi). This approach is called as unsupervised as we lack a response variable that can supervise our analysis. So, in this approach, we are in some sense working blind.

Difference between Supervised Learning and Unsupervised Learning
Unlike unsupervised learning, in supervised learning we have input predictor and response values to our model. In supervised learning we work on labelled data and in unsupervised we do not have labels to our data. In supervised learning approach, as name indicates model make prediction based on the learning it has from training data. In unsupervised learning, model builds a relationship among the data and the predictors to draw useful insights. As there is no response variable, entire data is used to build a model using unsupervised learning approach. But in supervised learning data used to divide into training and testing samples.

Commercial applications –

1. Supervised Learning –
   a. Predicting a salary of an employee based on his/her experience and job position by referring salary data of all employees of that company.
   b. Predicting weather forecast based on historical (previous weather reports) data.

2. Unsupervised Learning -

a. Categorizing the given defects of a certain product based on issue area/problem domains learned from the detail defect description analysis.
b. Image Segmentation

Question 2
Supervised Learning: Regression versus Classification [7 points]: In your own
words, define each learning problem, describe the difference between them, and name
two commercial applications learned using each approach (i.e., 4 applications in total).

Answer –
Variables of data can be categorized as either Quantitative or Qualitative/ Categorical. Quantitative variables deal with numerical values. Ex. – Gross profit, population of a country. In contrast, qualitative variables classify values in a one of the available classes. Ex. - Country Name, Size of a Firm etc.
So, in most cases problems with quantitative response as regression problems. And those involved qualitative responses are often refered to Classification problems.

So, main distinction point between Regression and Classification deals with what is the aim of your solution, if you want to categorize the values into available classes then it's a classification problem. But if you want to predict something based on available training numerical data then it's a regression problem.

Difference between Regression and Classification –
In regression, we predict the response value based on training samples of predictor response value pairs. In classification, we categorize each observation into available 'K' classes based on observation and class characteristics comparison. Regression deals with quantitative variables, Classification deals with Qualitative/Categorical variables. Predicting a population of a country in a X year based on historical population data is an example of a regression setting. From large available set of toys, classifying the toys based on their shape (ex. Rectangle, circle, triangle) is a classification problem.

Commercial applications of Regression and Classification –

1. Regression
   a. Predicting export turnover of a company/firm based on available financial data of different firms established in a particular region (city/ country).
   b. Predicting stock market/ share prices/ cryptocurrency prices based on tweeter discussions/ historical currency value data.
2. Classification -
   a. Predict whether firm can export or not based on export turnover and sales values of other firms present in training data.
   b. Classifying a firm into one of the four categories (very small/ small/ large/ very large) based on available training financial data along with size categorization of all the firms available in training data.

Question 3
Supervised Learning: Generalization [5 points]:
(a) Describe the motivation for splitting a dataset into a training partition and a test partition.
(b) If your model performs well on the training data and generalizes poorly to new instances, is the model overfitting or underfitting?
(c) If your model performs poorly on the training data and poorly on new instances, is the model overfitting or underfitting?

Answer –

a. If we develop some predictive models then we must have a way to assess their accuracy, reliability. So, assessing the quality of a model means we would like to know what will happen if we use this model in making real time prediction on unknown data. Will our predictions be relatively close to the actual outcomes? So, simplest way to handle this is by simulating creation of future data i.e by reserving some part of available data to be consider as future data. So, we usually divide the data into training and testing data. Model is built on training data and test data can be used as a simulated version of future data to find prediction and then to compare those prediction with the actual values of test data to define the accuracy and reliability of a model.

b. It is a case of Overfitting as model is purely shaped as per the training data but not at all working good for future data/test data.

c. It is a case of underfitting as model is not fit for both training data and future/test data. This might be a case of wrong method selection.

Question 4:
Offline versus Online Learning [6 points]: In your own words, define each learning approach describe the difference between them, and describe at least one advantage for using each approach.

Answer-
Offline Learning – The learning where the data set which we have is static. There is no updation in it. Analysis is done based on data set present.
Online Learning – The learning where the data is changing which means it is dynamic. Analysis can be done on the very recent data set, which can be further modified or updated and added.

Advantage:

Offline – Offline model is the only option in many of the businesses. Eg. A Café shop can advertise or display the menu card on social media, but obviously cannot serve online. Hence, offline data is effective. Offline dataset is may businesses.

Online – Updated data set, bulk data which can help in good research.

Differences between Offline and Online Learning:

Offline	**Online**
1. Data set modification is not possible in real time. 2. Fails to provide latest statistics. 3. Batch jobs which processes the data keeps on running which takes time. It becomes more tedious if even one of them fails to run.	1. Data set can be modified or updated in real time. 2. Updated latest statistics or numbers can be obtained. 3. Low Latency in processing the data set.