EndTerm Report

This project was a really new experience for me as I have never studied or worked with anything related to ML or AI before. This project really ~~gave a~~ has given me a foundation to proceed further.

STATISTICAL LEARNING :-

→ Supervised Learning

Every ~~inp~~ training dataset fed to the model has the label of the output along with the input, then it trains to map the input to the output and predict labels for a different test dataset.
Under this, linear regression, logistic regression, classification and tree methods were things which I learnt in detail.

→ Unsupervised learning

No labelled dataset is given to the model and it itself tries to understand the patterns in the data by clustering.

LINEAR REGRESSION :-

This is the basic algorithm in which the prediction is expressed as a linear combination of several data features of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n$$

Throughout this chapter I learnt how to represent situations as linear functions. Quantitative data can be directly represented and qualitative data can be represented by setting a dummy variable which can be 0 (no relation) or 1 (relation exists).

Then the method of estimating the coefficients using least squares regression was covered. The coefficient could be found by minimising RSS, and ~~analy~~ the relationship could be analyzed by various parameters such as R^2 statistic, t-statistic, p-value, etc. And which coefficients contribute a lot to the output which do not solely, but create a significant relation when correlated with other features.

CLASSIFICATION :-

Learnt when to apply classification techniques and when to apply regression techniques.

Logistic regression is the most commonly used technique for classification, in which the ~~th~~ logarith of the likelihood function has a linear relationship.

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x_1+\ldots} \Rightarrow \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x_1+$$

A significant difference between linear and logistic regression is what they predict. Linear regression predicts the value of some quantity in terms of certain features. whereas logistic regression predicts the conditional probability of a certain class given some other condition is happening.

Then came other methods of classification (generative methods) :

Linear Discriminant Analysis - Related to a probability density function based on the mean and variance of the classes, and how the probability can be maximized and coefficients estimated accordingly.

Quadratic Discriminant Analysis - A complex development to LDA, a lot more confusing.

Naive Bayes - An important technique of predicting posterior probability using Bayes' theorem

K- Nearest Neighbours -

Setting some k- value and finding the decision boundary that sets up observations to their class. This is done by finding the k no. of closest observations to a particular chosen observation and estimates the conditional probability for classes and sets the observation to the class with maximum probability. More the k value, lower the flexibility of the decision boundary resulting in high bias and low variance.

Confusion Matrix is something which was a revision of my previous knowledge, which is a way to depict how the model predicts true and false classifications correctly. Some metrics associated with confusion matrix are sensitivity, specificity, accuracy and F1 score.

ERRORS :-

Some of the different types and methods of error estimation covered through the phase of this project were -

- Mean squared error (MSE) : Average of squared difference between actual and predicted value.

- Standard error : ~~Vo~~ Standard deviation, root of variance

- t-statistic : Determines if the null hypothesis in linear regression can be ignored or not. ~~↑ t-statistic~~ More the p-value of a t-statistic, more the null hypothesis is true.

R^2 statistic: Ratio of RSS, btw 0 and 1

TREE - METHODS: -

Decision Trees are easily explainable approaches to classification and regression.

Regression Trees narrow down certain ~~yes/no~~ yes/no condition starting from a root. The most important parameter is kept at the first and further divisions occur at each branch. Visually it is like dividing the whole region ~~cond~~ containing the datapoints into various sub-regions at each step.

Error rate for a decision tree can be measured using cross validation. In k-fold cross validation, the observations are divided into k folds, recursive splitting is done till there are less than some minimum number of observations ~~in~~ in each terminal ~~done~~. node.

classification trees are similar to regression trees but the final leaf node is a label (yes/no) rather than a quantitative value.

Bagging : essentially it is making a prediction using different datasets as training and testing data, and averaging them all over. In ~~regr~~ decision trees it is like having different ~~regression~~decision trees using different bootstrapped observations and averaging them to reduce variance.

Random Forests : Random Forests is similar to bagging, but at each split, a random sample of m predictors is taken out of the p total predictors, and this $m \approx \sqrt{p}$. The main advantage of this over bagging is that the many decision treated in bagging have high changes of being correlated which increases the variance. But this is not the case in random forests as all chosen predictors are random.

NEURAL NETWORKS:-

This is a completely new and very different learning method, than the other supervised learning methods till now. It involves deep learning and typically consists of an input layer, ~~many~~ many hidden layers and an output layer, having many non-linear functions going into different layers.

ReLU function is used here unlike sigmoid which makes the -ve values go to 0.

Convolution Neural Networks ~~is~~ are widely used in image classification consisting of consecutive convultion and pooling layers. Convolution applies horizontal and vertical filters to extract small elements from the image and pooling converts larger to smaller image.

In a nutshell, I really learnt a lot from this project and hope to use this basic understanding in the upcoming ones :)