Assignment Supervised Learning

Q1. Which technique(s) from the following would you use?

- **Logistic Regression**
- **Support Vector Machines**
- **Bayesian classifier**
- **K-Nearest Neighbour**
- **Decision Trees.**
- **Principal Components Analysis**

1. **Care only about accuracy and not about interpretation?**
    a. I used Support Vector Machines (SVM) for more accuracy, because SVM uses the kernel technique which transforms the data and finds optimal barrier between the likely outputs for smaller data sets.
    b. If there is huge data, It is better to use K-Nearest Neighbor or Decision trees which are faster and also gives better accuracy than other classifiers.

2. **Care about interpretation more than accuracy?**
    a. K- Nearest Neighbour is a best known for ease to interpret the output. The scatter plot from knn itself shows how the boundary is shared between the data points, which also help to understand how the data is dividing with an increase of k value(higher the k smoother the curve). The plots with the validation error assistant to find k value easily. Hence KNN is more beneficial to use.

3. **Scoring throughput is high?**
    a. When the input is high it is more beneficial to use either KNN or Bayesian classifier for classification. To get a better score use KNN, it performs more reliable than Bayesian classifier

4. **Want to build a robust and deterministic model?**
    a. It is more suitable to use Logical Regression because it gives the relation between the response variable and regression variables. Further, it is known for better training speed and predicting speed.

5. **Each class has potentially multiple sub-classes?**
    a. SVM and decision tree algorithm are better for more subclasses..

6. **Want to visualize data that has no class labels?**
    a. We use unsupervised technique for label data. So, we use Principal Components Analysis (PCA).

7. **Don't want to build a model at all?**
    a. K- nearest neighbour Non-parametric classifier. Predictions are made calculating the similarity between the two input and training instance. So we need not build the model at all.

Q2. Five types of problems:

1. ClassificationProblem
2. RegressionProblem
3. Retrieval Problem
4. Recommendation Problem
5. Reasoning Problem

Identify and explain that below domains belong to which class of problems (any 2 areas)

1. **News.google.com**
    a. It uses **Classification** for searching, which classifies the web pages and the information. Further, it retrieves the information from the user, along with that information, it relates the user history and **recommends** the news for the user.

2. **Youtube.com**
    a. Similarly like google news it retrieves the data from the user send the information to the database, which **classify** the videos for particular category and gives videos based on that. It uses the user history to **recommend** the videos.

3. **Linkedin.com**
    a. Suggestions on Linkedin are managed based on **recommendations** using contacts or the information related to yourself (such as bio.). It also uses **retrieved** data and shows the suggestion profiles of people.

4. **ola.com/uber.com**
    a. Ola and Uber use **regression** for calculating the fare for the ride. This is because regression helps to predict continuous values. It also **recommends** the best available route as per the fare and type of vehicle.

5. **99acres.com / magicbricks.com**
    a. 99 acres and magic bricks work on bases of **recommendations** and classification. It classifies the available houses, based on that query. Further, based on search history and location it will **recommend** the houses.

Q3. List one difference and one similarity between the following:

1) K-nearest Neighbour and K-Means

a) **Difference:** K- Means is Clustering Technique whereas K-NN is classification technique
b) **Similarity**: Both work for the labeled data.

2) PCA and MDS

a) **Difference:** Pca used to determine the correlation among samples whereas MDS is used to calculate the distance among the samples.
b) **Similarity :** Both are belonged to feature extraction methods.

3) Classification and Regression

a) **Difference:** The response variable for classification is numerical (continuous) whereas for classification is categorical (discrete).
b) **Similarity:** Both are under the umbrella of supervised learning.

4) Bag-of-words and Market Basket Analysis

a) **Difference :** Bag of words is used in Natural Language Processing and information retrieval whereas Market Basket analysis used to perform build the association rules which are helpful to sell the products.
b) **Similarity:** Both features are collection of set of things.

5) **Naïve Bayes and Decision Trees**
    a) **Difference :** For the Naive Bayes you need to pick the features whereas for Decision trees it will pick the best features from the tabular data.
    b) **Similarity:** Both are used to classify the data.

Q4. In this problem, we will explore what dataset and within those datasets what features you want to create for solving the following problems. [Give at least four features for each problem]

1) **You are a new Fintech startup in town. You Build Credit Models Based on people's SMS data (financial related SMSes only). What kind of features will you engineer that will help you predict the credit score of your customers?**

    a) I will consider at least 6 features before building the model for predicting the score credit score. The Basic six are:
        i) **Monthly Income (numeric)**
        ii) **Monthly Expenditure (numeric)**
        iii) **Monthly Maintenance of credit.(good balance) (numeric)**
        iv) **Payment of Installments in time (numeric i.e, discrete)**
        v) **Delay of Installments. (numeric i.e, discrete)**
        vi) **Past installments rating (numeric, i.e, ratings**)

    b) Based on the above we can build a strong model to predict the accurate model using Linear regression. Each factor is independent of one another and each one will be available from the messages. Variation in income may help to enhance the credit score and assist to increase the new credit limit. Expenditure tells how great they are spending monthly. Maintenance of good credit assistant to increase the understanding of security. Payments and delay will improve and minimise the credit score sequentially. Past instalments will also impact a lot to predict the credit score.

2) **Let's say WhatsApp hires you to build the following model. WhatsApp wants to find the most "popular" people in its customer base for viral marketing. What features on each person you will you use to come up with a popularity score of a person on WhatsApp. (Note: A person might be part of multiple groups)**

a) I will consider five features building the model, where 3 features are divided into subclasses.

- **i) Active in Whatsapp:**
  - **(1) Group**
  - **(2) Individual**
- **ii) Number of income msgs:**
  - **(1) Group**
  - **(2) Individual**
- **iii) Number of outgoing msgs:**
  - **(1) Group**
  - **(2) Individual**
- **iv) Coverage of contacts. (In how many phones he got covered, In his phone and others too.)**
- **v) How well people are responding to his messages.**

b) The above mentioned are properties assists to build a model that will benefit WhatsApp to discover the person for viral marketing. The active feature will explain to us how he/she is utilising the platform and also how properly he is responding to messages in groups and as well as a person. The number of income messages will improve to determine the level of his/her popularity. The number of outgoing messages and the response to his/her will progress or diminish the popularity of him/her. The coverage indicates his popularity directly. The more the coverage the higher he is spread. Then there is a higher chance of spreading the information.

K.V.MURALI KRISHNA