Assignment 1

Problem 1 (3 marks)

Select your 10 favourite TV series (atleast two of them ongoing) and your 10 favourite movies. Plot the total time duration of each of the entries in both the cases in separate figures. For a TV series, the total time duration(T) is given as T = approx time of an episode * no. of episodes. Please submit the csv files and the plots.

Problem 2 (3 marks)

Suppose you need to organise a group study program in your neighbourhood. In order to organise it, collect the data from 20 people from your neighbourhood about the time of the day when people are generally most focussed.(It could be anytime from 0000 to 2359 hrs). Make an appropriate plot (search for it if you don't know of any) to show the distribution of the time and comment on what is your suggested time of the group study program. Please submit the csv file having only one column showing the preferred timing of the selected 20 people. Please submit the plot.

Problem 3 (3 marks)

Let Y be a categorical response variable taking values in $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ and X be the set of covariates. Under the zero-one loss function, show that the expected loss is minimised for any given $X = x$, when the predicted category is given by $\hat{G(x)} = \max_g Pr(Y = g|X = x)$, where $g \in \mathcal{G}$.

Problem 4 (4 marks)

Let $Y = f(X)$ where $f(X) = \exp(-8||X||^2)$ and $||X||$ denotes the norm of any vector $X \in R^p$. Generate 1000 points uniformly on the cube given by $[-1, 1]^p$ for $p \in \{1, 2, \ldots, 10\}$. Predict the value of the function at $X = (0, 0, \ldots, 0)$ using the 1-nearest neighbour for each p. Plot the squared error distance between the true value i.e. $f(0)$ and the estimated value based on 1-NN with the dimension. (squared error distance on y-axis vs dimension on the x-axis). Comment on the phenomenon observed based on this plot of dimension vs squared error distance.

Problem 5 (5 marks)

Let Y be the response variable taking values in 2 categories "blue" and "red"; and X be the covariate. Generate 100 observations for $X \in R^2$ from a mixture of bivariate normal

distributions with means $(3, 0)$ and $(0, 3)$ and Covariance Matrix as Identity with mixture proportions 0.8 and 0.2, respectively. Categorise all the responses in this case as "blue". Next, generate $X \in R^2$ from a mixture of bivariate normal distributions with means $(0, 3)$ and $(3, 0)$ and Covariance Matrix as Identity with mixture proportions 0.8 and 0.2, respectively. Categorise all the responses in this case as "red". Take both these set of responses and covariates as training set. Plot the training set observations with different colours for different categories. Apply the k-nearest neighbour to estimate the category of the response on the training set considering the neighbours from the training set. Choose k=1, 10, 100, 150, 200. Plot the average zero-one-loss against k.

Repeat the whole procedure again but take 500 samples from both the categories and denote the generated observations as test dataset. Apply the k-nearest neighbour to estimate the category of the response on the test dataset considering the neighbours from the training set. Choose k=1, 10, 100, 150 ,200. Plot the average zero-one-loss against k.

Problem 6 (3+3 marks)

a) In the dataset "data6.csv", regress y on x1, x2, x3, *cat_variable* and x7. Report the estimates, test the hypothesis for significance of each variable as well as the whole model, confidence intervals, p-values, residual plots. Which variables are statistically significant ?
b) In the same dataset, perform forward stepwise regression and choose the appropriate variables. Show the whole stepwise procedure.

(Here, the *cat_variable* is a categorical variable with three levels).

Problem 7 (3 marks)

In the dataset "data7.csv", fit $y = \beta_0 + \sum_{j=1}^{p} \beta_j (x)^j + \epsilon$. Choose the appropriate p and show the coefficients along with the p-values.

Problem 8 (3 marks)

In the dataset "data8.csv", regress y3 on V2 and V3. Identify the outliers and the leverage points.