Theoretical Assignment 2

Due Date: 18th November — Total Marks: 30

Q1. Gradient-boosted Regression

Gradient Boosting is a more advanced version of Adaboost that has become popular recently. It is used mainly for regression. Implement Gradient-boosted linear regression as follows:

i) Create a dataset of 10 points, by drawing 10 random numbers and scaling them to the range (-50, 50). *Use the last 5 digits of your roll number as random number seed*. These are the values of 'X'. For each 'X', generate 'Y' according to the rule $Y= X^2+7x+4$. Use first 8 points for training, and rest 2 for testing.

ii) Calculate the mean of all the output values Y in training set. Save it as Y0. Define residuals for each datapoint: Ri = Yi – Y0.

iii) Carry out linear regression to predict Ri from Xi. Save the linear regression coefficients as (W1, b1).

iv) Predict output variables as Y1i = Y0 + α1*(W1*xi + b1). You should choose α1 such that $\sum_i(Y1i-Yi)^2$ is least. Alternatively, you can choose α1 as a small constant below 1.
Calculate the new residuals Ri = Yi – Y1i.

v) Repeat steps iii) - iv) for 2 more iterations. Save the regression coefficients as (W2, b2), (W3, b3). Estimate Y2i = Y0 + α1*(W1*xi + b1) + α2*(W2*xi + b2), and Y3i = Y0 + α1*(W1*xi + b1) + α2*(W2*xi + b2) + α3*(W3*xi + b3).

vi) Make predictions on the two test datapoints in the same way, using the (α, W, b) as computed already.

Reference: https://towardsdatascience.com/machine-learning-part-18-boosting-algorithms-gradient-boosting-in-python-ef5ae6965be4 (though the example here is for decision trees)

Q2. K-means++ Clustering

With a small dataset of N 2D points having K "natural clusters", illustrate how K-means may end up creating "wrong" clusters due to poor choice of initial clusters. Also show how K-means++ helps to improve the situation.

Take K as the *last digit of your roll number, plus 2.* Choose N accordingly, such that each "natural cluster" has at least 3 points. **Choose the natural clusters in your own way, but make sure that there is some geometric basis to consider them as natural clusters.**

For both K-means and K-means++, show only ceiling(11/K) number of iterations. **[5+5=10 marks]**

Q3. Parameter Estimation

Derive maximum-likelihood estimates of the parameters of i) Geometric distribution ii) Binomial Distribution. Also derive the posterior distribution of the variance parameter of 1-dimensional Gaussian distribution, assuming that the mean parameter is known. Use “inverse-gamma” as prior.

[2+2+6 = 10 marks]