MTCS-205 Assignment 1: Chapters 1-4

Welcome to the first assignment in the MTCS-205 lab!

Prerequisites: chapters 1-4.

Skills to be tested: Python coding ability
Knowledge of prerequisites

Assignments:

1. Generate a sample set of size N from the Gaussian distribution with mean 0, variance 1. Deriving the equations for maximum likelihood estimation, compute the estimated mean and variance. Also, compute the unbiased estimators for the mean and the variance.
Perform this experiment for 100 runs and tabulate the following results for each value of N = {10,100,1000,10000,100000}:
a. Difference between the actual mean and the mean given by the maximum likelihood estimator
b. Difference between the actual mean and the mean given by the unbiased estimator
c. Difference between the actual variance and the variance given by the maximum likelihood estimator
d. Difference between the actual variance and the variance given by the unbiased estimator.
Additionally, Compare the estimates made by the two methods. Comment on their behaviour.

2. Generate labelled datasets (both training and test) from 5 different Gaussian distributions. Details about the generation are given later. Using these datasets, make predictions for the test set using the posterior as the discriminant function. Tabulate classification accuracies on both the training set and the test set using the following assumptions:
a. Equal priors and variances
b. Estimated priors and equal variances
c. Estimated variances and equal priors
c. Estimated priors and variances

Details about the generation of the datasets:
The following are the means and variances of the

distributions
a. mean = 0, variance = 1
b. mean = 10, variance = 2
c. mean = 5, variance = 4
d. mean = 20, variance = 1.2
e. mean = 15, variance = 7

Generate, once for a condition, a ‘training’ set of size 50000 and a ‘test’ set of size 10000 using the following conditions:

a. Each distribution is equally likely (This is not the same as equal number of samples)
b. Use the following priors: P(C_1) = 0.1, P(C_2) = 0.25, P(C_3) = 0.3, P(C_4) = 0.25, P(C_5) = 0.1

The labels for the data are between 0 and 4 depending on the gaussian that generated them.

3. Perform association rule mining using the apriori algorithm for the data at this link: https://drive.google.com/file/d/1y5DYn0dGoSbC22xowBq2d4po6h1JxcTQ/view?usp=sharing

Abstain from using library functions. Use a minimum support of 0.3 and a minimum confidence of 0.8. The associations rules should be tabulated.