**Statistical Methods in Artificial Intelligence – Spring 2025**

Quiz-1

Answer Key

Q1) (1 mark)

Correct Ans: Training Error: Increases, Testing Error: Decreases

Explaination should be along the lines of: (if explaination subpar, only partial marksa re provided)

1. As training set becomes larger, it becomes difficult to fit

2. With larger set, training improves and it generalises. So test error reduces

0.25 for each correct answer and 0.25 for each explaination. If the answers are wrong, explainations will not be considered/

Q2) (1 mark)

0.5 marks each for part a, b. Further breakup- 0.25 for using MSE 0.25 for using partial differentiation to find the parameter values correctly

Q3) (1 mark)

0.5 marks: for writing the correct formulas for both cases- Tf\*IDF for doc-doc, Log(1+Tf)\*IDF for query-doc (per term)

0.5 marks: point out differences in the formulae i.e. use of Log in query to doc, normalized tfs in case of doc-doc but not in phase query-doc, reasons for change: 1 added to remove the effect of words not in phase query (it makes the Log term 0 as Log (1)=0), very few words in query so normalization using total terms is not used as it might skew the score (includes count of terms which are not present in query), Log acts as its replacement Q4) (1 marks)

1/3 mark for each.

2 major challenges that have to be there are High computational cost (use efficient nearest neighbor search) and Curse of Dimensionality (use PCA to reduce dimensionality).

3rd can be Imbalanced Dataset (use weighted KNN), choosing optimal K- value (use cross validation), Data Representation (normalize using min max scaling or standardization), choice of Distance metric (choose according to the type of data type), Sensitivity to noise and outliers (Distance weighted KNN and outlier removal).

Q5) (2 marks)

1 mark for each sub part (a) and (b) Part a => 0.5 for the correctness of the final decision tree, 0.25 for showing calculations for entropy/ Information gain, 0.25 for Truth Table. Same marks distribution for part b also.

Q6) ( 1 Mark)

Outlier point gets assigned as the center or k > n -> 1 mark kmeans++ mentioned and then written that it cant be empty -> 0.5 marks