CS685A: Data Mining

Endsem: Part 1
Total Marks: 50

13th December, 2020: 9:15–10:15 am

Instructions

This question paper contains **15** questions in **2** pages.

Multiple choice questions carry a negative marking of 1 mark for wrong answers.

Q1: [1 mark] Leave-one-out cross-validation technique is a k-fold cross-validation method with $k = 1$.

A. True B. False

Q2: [1 mark] The following formula is correct: prior $\propto$ (evidence $\times$ posterior)

A. True B. False

Q3: [1 mark] An iceberg cube is one where the number of dimensions aggregated is above a threshold.

A. True B. False

Q4: [1 mark] The assumption for naïve Bayes classifiers is that the attributes are independent.

A. True B. False

Q5: [1 mark] Given enough training time, an ANN can always model a continuous function correctly using only one layer of hidden nodes up to any arbitrary error factor.

A. True B. False

Q6: [1 mark] If two classes are linearly separable, then SVM learns a classifier that is correct for all training objects.

A. True B. False

Q7: [1 mark] For a real symmetric matrix, singular values and eigen values are the same.

A. True B. False

Q8: [1 mark] Mahalanobis distance takes into account only the covariances but not the variances.

A. True B. False

Q9: [1 mark] The testing of lesser than operation cannot be done on ordinal data.

A. True B. False

Q10: [1 mark] CLARA runs PAM on a sampled dataset.

A. True B. False

Q11: [8 marks] Consider the set of training objects for mangoes (M) and oranges (O) using attributes color and size.

(M,yellow,large); (M,yellow,medium); (M,yellow,large); (M,yellow,large); (M,yellow,small); (M,yellow,large); (M,green,medium); (M,green,large);

(O,orange,medium); (O,orange,small); (O,orange,medium); (O,orange,medium); (O,green,small); (O,yellow,medium);

Use a naive Bayes classifier to state the class of the following fruit: (yellow,medium)? (Use a Laplacian correction of $\epsilon = 0.05$.)

Q12: [8 marks] Consider a decision tree binary split using Gini index.

Suppose the original dataset D contains 60 objects of class A, 30 of class B, and 10 of class C. After the split, the two parts contain the following distribution of objects from the three classes A, B, C respectively:
D_1: 45, 25, 5; D_2: 15, 5, 5.

Find the Gini index of D and the split. What is the gain?

Q13: [8 marks] Assume a dataset of 10 fruits in which the first 4 ($O_1 - O_4$) are mangoes. A classification algorithm returns O_1, O_2, O_4, O_5, O_8 as mangoes.

Find the precision, recall and F-score of the algorithm with $\beta = 1$, 2, and 1/2.

Q14: [8 marks] Suppose there is a roll of a fair dice. An i^{th} roll is called a *hit* if it shows the value i. Each experiment consists of 6 consecutive rolls of the dice. An exeriment is called a *success* if there is at least 1 hit in it. What is the expected chance of an experiment having success?

If after 10 such experiments, 9 successes are recorded, what is the chi-square value of it (assume a one-sided tail)?

Q15: [8 marks] Consider the following set of points and their coordinates.

$A : (12, 23)$, $B : (89, 27)$, $C : (42, 42)$, $D : (63, 83)$, $E : (95, 10)$.

Perform a hierarchical clustering using average pairwise distance.

Show clearly the steps and the final dendogram.