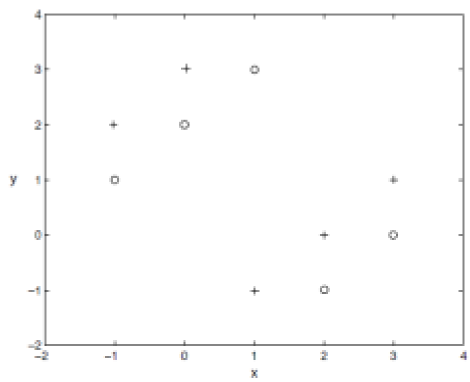
Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2015-2016
CS F415 / IS F415: Data Mining
Comprehensive Examination – Part B

Type: Closed **Time: 180 mins** **Max Marks: 58** **Date: 12.05.2016**

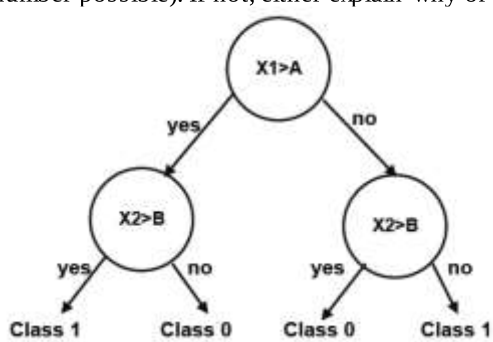
All parts of the same question should be answered together.

1. Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and o). [2 + 4 Marks]



a. What is the error when using 1-NN?
b. Which of the following values of k leads to the minimum number of validation errors: 3, 5 or 9? What is the error for that k?

2. Assume we have the decision tree in the below mentioned Figure which classifies two dimensional vectors {X1, X2} ∈ R \ {A, B}. In other words, the values A and B are never used in the inputs. Can this decision tree be implemented as a 1-NN? If so, explicitly say what are the values you use for the 1-NN (you should use the minimal number possible). If not, either explain why or provide a counter example. [6 Marks]



3. In linear PCA, the covariance matrix of the data $C = X^TX$ is decomposed into weighted sums of its eigenvalues (λ) and eigenvectors p:

$$\mathbf{C} = \sum_i \lambda_i \mathbf{p}_i \mathbf{p}_i^T$$

Prove mathematically that the first eigenvalue λ_1 is identical to the variance obtained by projecting data into the first principal component p_1 (hint: PCA maximizes variance by projecting data onto its principal components).

[6 Marks]

4. The E-step in estimating a GMM infers the probabilistic membership of each data point in each component Z: $P(Z_j / X_i)$; $i = 1,\ldots,n$; $j = 1, \ldots, k$, where i indexes data and j indexes components. Suppose a GMM has two components with known variance and an equal prior distribution

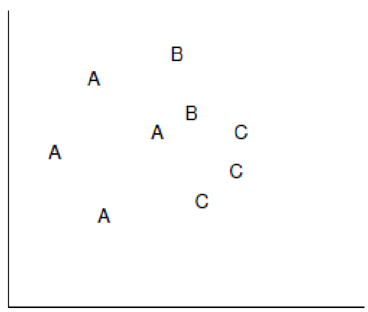
$$N(\mu_1, 1) \times 0.5 + N(\mu_2, 1) \times 0.5$$

The observed data are x1 = 2, and the current estimates of μ_1 and μ_2 are 2 and 1 respectively. Compute the component memberships of this observed data point for the next E-step (hint: normal densities for standardized variable $y_{(\mu=0;\ \sigma=1)}$ at 0, 0.5, 1, 1.5, 2 are 0.4, 0.35, 0.24, 0.13, 0.05). [8 Marks]

5. Consider learning a target function of the form $f : R^2 \rightarrow \{A, B, C\}$ that is, a function with 3 discrete values defined over the 2-dimensional plane. Consider the following learning algorithms:

- Support Vector Machine
- 1-nearest neighbor

Note each of these algorithms can be used to learn our target function f, though doing so might require a common extension.



For each of these algorithms,
A. Describe any assumptions you are making about the variant of the algorithm you would use
B. Draw in the decision surface that would be learned given this training data (and describing any ambiguities in your decision surface)
C. Circle any examples that would be misclassified in a leave-one-out evaluation of this algorithm with this data. That is, if you were to repeatedly train on n-1 of these examples, and use the learned classifier to label the left out example, will it be misclassified? [8 Marks]

6. a. Assume we are trying to cluster the points $2^0, 2^1, 2^2, \ldots, 2^n$ (a total of n points where $n = 2^N$) using hierarchical clustering. If we are using Euclidian distance, draw a sketch of the hierarchical clustering tree we would obtain for each of the linkage methods (single, complete and average). [6 Marks]
Hint: All the 'n' points are on a real line.

b. Now assume we are using the following distance function: d(A;B) = max(A;B) / min(A;B). Which of the linkage methods above will result in a different tree from the one obtained in (6.a.) when using this distance function? If you think that one or more of these methods will result in a different tree, sketch the new tree as well. [6 Marks]

7. Human eyes are fast and effective at judging the quality of clustering methods for two-dimensional data. Can you design a data visualization method that may help humans visualize data clusters and judge the clustering quality for three-dimensional data? What about for even higher dimensional data? [6 Marks]

8. Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations. [6 Marks]
a. k-means
b. DBSCAN