CMSC 678 – Fall 2016 — Homework 1
RUJUTA PALANDE UMBC ID : OT30378

September 20, 2016

1 True or False

(a) true
(b) false
(c) Not sure
(d) true
(e) true
(f) true
(g) false
(h) true
(i) true
(j) not sure
(k) not sure

2 Go through matlab tutorial

3 In class, we looked at an example where all the attributes were binary (i.e., yes/no valued). Consider an example where instead of the attribute “Morning?”, we had an attribute “Time” which specifies when the class begins. (a) We can pick a threshold tau and use (Time ¡ tau)? as a criteria to split the data in two. Explain how you might pick the optimal value of tau . (b) In the decision tree learning algorithm discussed in class, once a binary attribute is used, the subtrees do not need to consider it. Explain why when there are continuous attributes this may not be the case

(a) The optimal value of the threshhold tau will be such that the test data will be distributed in two halves where similar kinds of test data fall in 2 halves. Thus the data will be split in two.
(b) In continuous attributes there are many possible outcomes unlike decision tree algorithm with binary attributes.So in case of continuous attributes, we need to consider the subtrees as well.

4 Why memorizing the training data and doing table lookups is a bad strategy for learning? How do we prevent that in decision trees?

If the algorithm has memorized the training data and doing table lookups, then the algorithm is not generalized and hasn’t learnt anything from the training data to apply on the test data. It won’t perform well on the unknown data or the test data. We prevent that in decision trees by using the pruning technique. We take each featre and divide the data based on feature. The unnecesarry features are not considered and thus memorization is prevented.

5 What does the decision boundary of 1-nearest neighbor classifier for 2 points (one positive, one negative) look like?

Consider ’A’ which represents the positive point and ’B’ which represents the negative point.The decision boundary of 1-nearest neighbor will be a line that is equidistant from both these points and passes through the line joining these two points. Basically the 2 points should be on the opposite sides of the decision boundary

6 Does the accuracy of a kNN classifier using the Euclidean distance change if you (a) translate the data (b) scale the data (i.e., multiply the all the points by a constant), or (c) rotate the data? Explain. Answer the same for a kNN classifier using Manhattan distance

For Euclidean distance Translate : Accuracy will remain the same since even if the points are translated, the distance between them won't change Scale : Accuracy will remain same for the similar reason as above Rotate : Accuracy will remain same for the similar reason as above For Manhattan distance, the accuracy varies for each of the three cases since the absolute difference between the co-ordinates changes