**April 5, 2016 Review for COSC 4355 Midtem2 Exam**

**1) DBSCAN**

**a)** Assume I run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and I obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=5 and epsilon=0.1. How do expect the clustering results to change?

**There number of core points will usually increase and therefore the number of outliers will t usually decrease (under) 5%. New clusters might occur and multiple clusters might be merged into a single cluster; consequently, it is not clear if the number of clusters will increase, decrease, or remain the same.**

**b)** Assume we have two core points o and v that are within each other’s radius—will o and v belong to the same cluster? Now assume, that we have a border point b within a radius of a core point u—will b and u always belong to the same cluster? Give reasons for you answer! [3]

Yes, o is density reachable from v and v is density reachable from o.

No if b is also in the radius of another corepoint w, and w is not density reachable from u; in this case b might end up in the cluster formed by growing around w, if we is processed before u.

**2) Classification**

1. Compute the GINI-gain[[1]](#footnote-1) for the following decision tree split (just giving the formula is fine!)[3]:

(12,4,6) (3,3,0)

(9,1,0)

(0,0,6)

G(6/11,2/11,3/11) – (6/22\*G(0.5,0.5,0) + 10/22\* G(0.9,0,1,0) + 0)=

(1- (6/11)\*\*2-(3/11)\*\*2-(2/11\*\*2)- (6/22\*0.5)- 10/22\*(1-0.9\*\*2-0.1\*\*2)=

(121-36-9-4)/121 - …=

72/121-,,,=

0.595-=

1. Assume there are 5 classes; Compute the entropy of the following class distribution: (1/2,1/4.1/8,1/8,0), giving the exact number not only the formula! [2]

H(1/2,1/4,1/8, 1/8,0)= ½\*log2(2)+ \*1/4log2(4)+ 2\*1/8log2(8)+0=0.5+0.5+6/8=1.75

c) Compare Nearest Neighbor Classifiers and Support Vector Machines! What are the main differences between the two approaches? [6]

|  |  |
| --- | --- |
| Nearest Neighbor Classifiers | Support Vector Machines |
| Consider only the neighborhood to classify🡪local classifer | Usually, Maps objects into higher dimensional space to classify |
| Uses multiple decision boundaries that are edges pf convex polygons that can be computed using Vornoi tessalations | Uses a single, global decision boundary which is a usually high-dimensional |
| Lazy learner, so doesn’t build a model of the training data | Builds the model of the training data, and usually uses a kernel function to map the data in higher dimensions |
| No training cost | Very expensive to learn |
| Distance function is critical; scale of attributes may impact predictions | Kernel function is critical |
| Can handle multiple class output directly | Can handle only two class output. To handle multiple class outputs, several binary classifiers need to be learned to separate instance of class from rest of classes |
| Finds local maxima only | Finds global maxima |

**d)** The following dataset is given (depicted below) with A being a continuous attribute and GINI is used as the evaluation function. What root test would be generated by the decision tree induction algorithm? What is the gain (equation 4.6 page 160 textbook) of the root test you chose? Please justify your answer![6]

Root test: A >=

|  |  |
| --- | --- |
| A | Class |
| 0.22 | 0 |
| 0.22 | 0 |
| 0.31 | 0 |
| 0.33 | 1 |
| 0.33 | 1 |
| 0.41 | 0 |
| 0.41 | 1 |

**Possible slits**

**A≤0.22: (0,2); (3,2)**

**A≤0.31: (0,3); (3,1)**

**A≤0.33: (2,3); (1,1)**

as A≤0.31has a purity of 100%/75% which is much higher than the purity of the other splits, this split will be selected.

e)Most decision tree tools use gain ratio and not GINI or information gain in their decision tree induction algorithm. Why? [3]

Information gain does not consider model complexity in terms of how many additional nodes added to a tree, whereas gain ratio does!

1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)