**November 5, 2018 Review for COSC 4355 Midtem2 Exam**

**3) 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) What is overfitting? What are the characteristics of overfitting? What can be done in the context of decision trees to battle overfitting?

The training error is low, but the testing error is not optimal.

Prune decision trees, reducing their size; use a large dataset to learn the decision tree

**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!

4. **Computing Entropy using R [11]**

Write a function *H* in R[[2]](#footnote-2), whose input is a vector of class proportions of arbitrary length[[3]](#footnote-3) called v (v contains O and positive numbers whose sum is exactly one) and returns the entropy of for v; e.g.

v<-c(0.5, 0.25, 0.25, 0)

H(v)

*would return:* 0.5\*log2(2) + 2\*1/4\*log2(4) + 0=1.5

Remark: Values of 0 in the input vector do not make any contributions to the overall entropy—their contribution is 0; therefore, make sure when you write the code of the H function that you do not compute 0\*log2(0) as this will return NA[[4]](#footnote-4).

H <- function(v){

H <- 0

for(i in 1:length(v)){

if(v[[i]] != 0){

 H<- v[i]\*log2(1/v[i])+ H}

 }

 return(H)

}

**5) SVMs [9]**

a) What are the characteristics of hyperplanes that support vector machines learn from a training set? [3]

The hyperplane separate the examples of the 2 classes, such that the examples of one class are on one side of the hyperplane and the examples of the other class are on the other side of the hyperplane[1.5].

The obtained hyperplane has the widest margin[1]---the empty space that separates the examples of the two classes is maximized! [0.5]

b) The soft margin support vector machine solves the following optimization problem:



What does the first term minimize? Depict all non-zero ξi in the figure below! What is the advantage of the soft margin approach over the linear SVM approach? [5]

All other points

have **ξi** values

of 0!

width

width

The inverse width of the margin with respect to the class1 and class2 hyperplane [1]. Depict [2; 2 errors=0 points]. Can deal with classification problems in which the examples are not linearly separable[2].

c) Referring to the figure above, explain how examples are classified by SVMs! What is the relationship between ξi and example i being classified correctly? [4]

Examples which are above the straight line hyperplane belong to the round class, and example below the line belong to the square class [1.5]. An example will be classified correctly if ξi is less equal to half of the width of the hyperplane; the width w is the distance between the class1 and class2 hyperplane. [2.5].

d) Assume you use an ensemble approach. What properties should base classifiers have to obtain a highly accurate ensemble classifier? [3]

The classifiers should be diverse; that is, they make different errors. [2]

The classifier should be “somewhat” accurate; their accuracy should be above 50%. [1]

6) Neural Networks

How are activation functions used in neural network computations? What is neural network learning all about? Give a brief sketch how multi-layer neural networks learn models.

Activation functions are applied the to the linear input of a node to determine the node’s activation/value. Neural network learning tries to find weights that minimize the error in the neural network prediction for a training set. Neural network learning adjust weights example by example adjusting weights in the direction of the steepest negative gradient of the error function---reducing the training error for the example at hand. Errors for intermediate layer are computed from the errors of the next layer using a backpropagation algorithm.

**7) KNN**

a) k-NN employ a lazy approach to learning models from training examples. What does this mean? What disadvantages you see with k-NN’s lazy learning approach? [3+1 extra point}

No model is learnt and consequently no training is performed; the data themselves are the model. [1.5]; slow; major computations have to be performed at runtime when determining the class label of an example. [1.5]

b) What can be said about the number and shape of decision boundaries of a k-NN classifier? [3]

The shape consist of lines formed of edges of convex polygons of the Vonoroi tessellation[1]; in general, decision boundaries can be non-convex and can take many forms [1]; K-NN classifiers can learn classifiers with multiple—not a single decision boundary as SVMs do— decision boundaries[[5]](#footnote-5).

c) How does a 3-nearest neighbor classify an example?

It finds the closest 3 examples in the training set and based on the class labels of those 3 examples determines that class, usually by using simple majority vote.

1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)
2. You will need to write your own code; calling a function in an R-package which computes entropy will not get much credit! [↑](#footnote-ref-2)
3. You can use the length function to determine how many numbers v contains. [↑](#footnote-ref-3)
4. Moreover, in R, log (8,2) computes log2(8). [↑](#footnote-ref-4)
5. Its number is only bounded by the number of examples in the training set. [↑](#footnote-ref-5)