Midterm2 Exam

COSC 3337 Data Science I

November 6, 2019

Your Name:

Your student id:

Problem 1 --- Decision Trees and Classification in General [20]

Problem 2 --- Similarity Assessment [10]

Problem 3 --- More on Classification Techniques [19]

Problem 4 --- SVM [13]

:

**Grade:**

[](http://en.wikipedia.org/wiki/Tiger)

The exam is “open books and notes” but the use of computers is not allowed; you have 75 minutes to complete the exam. The exam will count approx. 15-16% towards the course grade.

**1) Tree Classification Models and Classification in General [20]**

1. Compute the information-gain[[1]](#footnote-1) for the following decision tree split for a dataset containing 3 classes (compute the exact value; just giving the formula will only obtain partial credit) [6]

(1,1,2) (0,1.1)

(1,0,1)

H(1/2,1/4,1/4)- 2\*0.5\*H(0,1/2,1/2)=1.5-1=0.5

Correct Formula, wrong computation 3.5 points

One Error at most 3.5 points

2 Errors at 0-1 points

b) What are the characteristics of overfitting? What can be done to deal with overfitting when learning decision tree models? [4]

The model is too complex[0.5]; the testing accuracy is not optimal[1], the training accuracy is high[0.5]

Reduce model complexity[1]; increase size of training set used [1]

c) Why is pruning important when using decision trees? What is the difference between pre-pruning and post pruning? [3]

To avoid overfitting[1] Pre-pruning uses “stronger” termination conditions for the decision tree induction algorithm to obtain smaller trees[1] Post-pruning grows a large decision tree and then reduces its size (usually relying on a validation set)

d) Describe how 3-fold cross validation determines the accuracy of a classification algorithm! [4]

No answer given

e) Let us assume we use decisions trees for a numerical dataset. What can be said about the characteristics of the decision boundaries decision trees use to separate the classes in numerical datasets? [2]

uses axis parallel decision boundaries

f) What is the purpose of a validation set? [1]

To select appropriate (hyper-)parameters for the chosen machine learning algorithm

**2. Similarity Assessment [10]**

Design a distance function to assess the similarity of graduate students; each graduate student is characterized by the following attributes[7.5]:

1. Ssn
2. qud (“*quality of undergraduate degree*”) which is ordinal attribute with values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’.
3. gpa (which is a real number with mean 2.7 standard deviation is 0.5, and maximum 4.0 and minimum 1.8)
4. gender is an nominal attribute taking values in {male, female}.

Assume that the attributes qud and gpa are of major importance and the attribute gender is of a minor importance when assessing the similarity between students. Using your distance function compute the distance between the following 2 students: c1=(111111111, ‘good’, 2.7, male) and c2=(222222222, ‘poor’, 3.7, female)[2.5]!

We convert the qud values ‘excellent’, ‘very good’, ‘good, ‘fair’, and ‘poor’ using a mapping function **φ** to 4:0 and then divide number of values the ordinal attribute takes minus 1; that is, by 4 in this case**;** finally we compute the distancesby applying the L-1 norm to the mapped values of the attribute qud.

Normalize gpa using Z-score and find distance by using L-1 norm

dgender(a,b):= if a=b then 0 else 1

Assign weights 1 to qud, 1 to gpa and 0.2 to gender attributes. We obtain:

**d(u,v) = (1\*|(u.gpa – v.gpa)/0.5| + 1\*|φ(u.qud) – φ(v.dud)|/4 + 0.2\*dgender(u.gender, v.gender))/2.2**

**No normalization (no division by the sum of the attribute weights): 1 point penalty for an otherwise correct solution; otherwise, 0.5.**

**One Error: at most 4.5 of 7.5 points**

**Two or more errors: 0-1.5 points**

For 2 students c1=(111111111, ‘good’, 2.7, male) and c2=(222222222, ‘poor’, 3.7, female) we obtain for their distance.

d(c1,c2)= (2+2/4+0.2)/2.2=2.7/2.2≈1.2

**One Error in the computations for the example: at most 1 point**

**3) More on Classification [19]**

a) Assume you use an ensemble approach. What properties should base classifiers have to obtain a highly accurate ensemble classifier? [2]

The base classifier needs to have an accuracy of higher than 50%[0.5]; the base classifiers have to be diverse in that they make different decisions leading to different errors[1.5].

b) Why does AdaBoost increase the weights of misclassified examples? [2]

**To learn a classifier that makes different kind of errors [2]**

Problem 3 continued

c) What is the purpose of the backpropagation algorithm? What does the back-propagated error depend on? [4]

To compute the error for nodes in the intermediate layer(s) [1]

**a. The learning rate [1], the error in the successor node[1], the activation of the node whose back-propagated error is computed [1]**

d) How do neural networks find the “best” set of weights for a training set? Limit you answer to at most seven sentences! [6]

e) k-nearest neighbor classifiers employ a “lazy” classification approach—please explain! What are the disadvantages of kNN’s lazy classification approach? [5]

It does not create a model [1]; the data serve as the model

As no model exists it is not possible to explain how a decision is made [2]

As all computations are done when example are classified and not earlier the approach is quite slow [2]

**4) SVM and Kernels [13]**

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

svn-equation

What does the first term minimize (be precise!)? What role does C play? How many examples are misclassified in the figure below! What is the advantage of the soft margin approach over the linear SVM approach? [6]



It minimizes the inverse[0.5] margin[1]. C determines the weight that is assigned to each of the two objectives of the optimization problem (maximizing the margin / minimizing the error) [1.5]

six[1]

It is able to deal with datasets for which the examples are not linearly separable [2]

b) Support vector machines are frequently successfully used in conjunction with kernels—how does this approach work[2]? Why do you believe this approach has frequently led to high accuracies of the obtained classifier?[3] [5]

The dataset is mapped to a higher dimensional space [1] and a hyperplane is found in the mapped space not the original space [1]

As the mapped space is higher dimensional there are more ways / more possible hyperplanes to separate the example of the two classes with no error or a low error. [3]

As kernel functions are usually non-linear, decision boundaries in the mapped space correspond to non-straight line decision boundaries in the original space, facilitating finding good hyperplanes [2]

At most 3 points for the second question. Other answers to the second question might deserve partial credit!

c) What does the term “kernel trick” refer to? [2]

Although a hyperplane is found in the mapped space, computation to determine the hyperplane are performed in the original space, leading to lower complexities for obtaining the hyperplane.

1. Entropy before (computed using the H-function) and after the split. [↑](#footnote-ref-1)