Solution Sketches Midterm Exam A

COSC 3337 *Data Science I*

October 14, 2021

Your Name:

Your student id:

Problem 1 --- Supervised Learning and Decision Trees [9]

Problem 2 --- Miscellaneous Questions [16]

Problem 3 --- Neural Networks [13]

Problem 4 --- Support Vector Machines [9]

Problem 5 --- EDA [12]

59:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 20% towards the course grade. The use of computers and cell phones is strictly prohibited.

Write you answers on the exam paper; if you need more space use back of exam paper!

1. **Decision Trees and Supervised Learning [9]**
2. Compute the GINI-gain[[1]](#footnote-1) for the following decision tree split[[2]](#footnote-2) (give the formula and compute the actual value as well)! [4]

 (12,4,4) (0,4,4)

( 6,0,0)

 (6,0,0)

Gini-Before: 1-0.6\*\*2-2\*\*0.2=1-0.36-2\*0.04=1-0.44=0.56

Gini-After: 0.4\*0.5+0+0=0.2

Gini-Gain: 0.26 If error you can give up to 2 points partial credit.

b) 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]

axis-parallel straight lines; if they say rectangular decision boundaries also give them 2 points.

c) The decision tree induction algorithm is a greedy algorithm. What does this mean? [3]

Makes locally optimal decisions[1] Does not backtrack[1]gets to the goal state/finds a solution quickly[1] Is not guaranteed to find the optimal solution[1] At most 3 points!

**2) Miscellaneous Questions**

a) What is the purpose of validation sets when learning classification and prediction models? What is the purpose of test sets when learning classification and prediction models? [2]

validation set: used to select hyper parameter during training

test set: used to assess the accuracy for classifies unseen examples in straining or used to estimate the generalization a

1. Assume we have a dataset with an attribute A with a mean value 8(μ=8) and standard deviation 2(σ=2). According to the 68–95–99.7 rule, what is the probability that a value of attribute A is between 6 and 10? [2]

0.68

Problem 2 continued

c) An unnamed scientist uses training and testsets for supervised learning that share 25% of the examples—25% of the examples in the training set belong to the corresponding test set. So you believe this is a good idea? Give reasons for your answer! [3]

No [1] The purpose of testsets is to estimate the generalization error/to assess how well the classifier work of unseen examples. As the approach uses examples used in training in the testset, the obtained test set accuracies are no longer a good estimate for the generalization error, as examples used in training are used to assess the testset accuracy. Comment: Using training set examples in test sets is even considered as cheating by the ML community.

d) K-Nearest Neighbor classifiers are called lazy classifies; what does this actually mean? What disadvantages you see with k-NN’s lazy learning approach? [3]

No model is learnt[1]; the dataset itself is the model[0.5]; as no model exist it cannot be inspected to understand what the classifier learnt[1.5]; somewhat slow classification performance due to the absence of a model[1] at most 3 points!

e) Assume you use a decision tree learning tool and you observe overfitting. What could you do to reduce overfitting? [2]

reduce model complexity [1] increase number of training examples [1]

f) A confusion Matrix of a classification model for distinguishing apples from oranges and mangos is given below:

What is the accuracy of the classification model; what is its precision for class mango? What is its recall for class mango? It is okay to represent you answers as fractions; e.g. 17/36! [4=1+1.5+1.5]

Accuracy: 10/36

Precision Mango: 1/6

Recall Mango: 1/13

No partial credit

\

**3) Neural Networks [12]**

a) Assume the following subset of a neural network is given:

 wA,B=0.2

 ΔB=0.4

 A B

Assume that B is an intermediate node of a neural network, the forward propagation activation values of nodes aA and aB are 0.5 and 0.8 and the current value of WA,B is 0.2; the associated error ΔB of node B that was computed by the back propagation algorithm is 0.4, the learning rate γ is assumed to be 0.5. First give the general weight update formula and then compute the new value of weight wA,B! [5]

Formula: wAB= wAB + γ\*aA\*ΔB [1]

where aA donotes the activation of node A; we receive as the new weight of WAB:

WAB=0.2 + 0.5\*0.5\*0.4=0.2+0.1=0.3 No Partial Credit.

b) In general, what factors influence the size of the weight increase/decrease of weight wA,B—the step size of the weight update? [3]

Let us assume a weight w of an edge that connect node A to node B is updated: A🡪B

The steps size of the weight update depends on

* 1. The learning rate [1]
	2. The activation on node A [1]
	3. The derivative of the activation function of node B for its linear input [0.5]
	4. The error in node [0.5]

If they say instead of c. and d. the associated error for node B this is also correct.

c)Take a look at the sub neural network consisting of nodes A, B, C, and D in the figure below; give a formula that computes the associated error ΔA for a node A. Assume the used activation function is g and its derivative is denoted by g’, and the activation of a node X is denoted by aX and the linear input of a node X is denoted by zX. First provide a general formula; then, replace general variables in the formula by their actual known values. [3]

 wA,B=0.5

 ΔB=0.4

 A B

 wC**,A**=1 wD,A=0.5

 C D

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.5\*0.4=g’(0.2+0.2)\*0.5\*0.4=g’(0.4)\*0.5\*0.4

If the formula is not expanded just 1 point. If other errors at most 1 points partial credit.

We assume that g(0.4)=0.2

Problem 3 continued

d)You use a neural network tools and the tool indicated that the training took 500 epochs. What does that mean? [2]

During the training the ML algorithm went through the examples of the training set 500 times.

**4) Support Vector Machines [9]**

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



What does the first term minimize (be precise!)? What is the purpose of C? How many examples are misclassified in the figure below! Finally, add arrows to all examples in the figure below, whose ξi values are positive---the length of the arrow should correspond with the value to the respective ξi ![5]

The first term minimizes the inverse margin [1; only 0.5 is they do not say inverse 0.5];

C is a hyper parameter which controls the importance of the error in relationship to the size of the inverse margin [1]; 6 [1]; depicting the arrow correctly [2; up to one point partial credit]



b) SVMs have been successfully used in conjunction with Kernels. How does this approach exactly work? [2]

The dataset is mapped into new attribute space using the respective kernel mapping function. A hyperplane is learnt for the mapped space and used as the classification model.

c) Assume we use SVMs in the conjunction with a dataset that uses numerical attributes A1, A2, A3 and the learnt SVM uses the hyper plane:

 A1\*2 + A2\*3 – A3\*4 + 11

Give a (training) example that exactly lies on this hyperplane! [2]

(0,3,5) that is, A1=0, A2=3 and A3=5.

Many other correct answers!

5) Data Analysis [12]

1. Interpret the supervised scatter plot depicting male and females with their respective height and weight. Characterize the distribution of males and female in the attribute space. Assess the difficulty of the classification problem of predicting gender based on weight and height. [7]

 

The instances of the male class are in the weight range 80-140 and height range 57-73; the female class instances are in the weight range 55-110 and height range 52-66. [2.5] Concerning classification difficulty, examples with low heights/low weights examples tend to be female and examples with high heights/weights examples tend to be male, but the classification problem is challenging in the 57-66 heightX80-110 weight rectangle but easy for example outside this rectangle. [4.5]

Many other solutions particularly assessing classification difficulty might deserve partial or rarely full credit.

b) Assume a boxplot has been created using the following R-code for an attribute x, containing the indicated 12 values:

> x<-c (15,4,2,2,8,10,12,12,12,12,26,29)

> boxplot(x)

What is the median for the attribute x? What is the IQR for the attribute x? What the position of the lower and higher whisker of the box plot created for attribute x? Are there any outliers? Assume that outliers are values which are 1.5 IQR above the upper box boundary or 1.5 IQR below the lower box boundary. [5]

Median=11 [1]

IQR=12-2=10 [1]

Total of 3 points for the information below; can give up to 2 points partial credit of only one error!

Lower whisker: -2

Upper whisker: 26

Outliers: -15, 29

1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)
2. There are 3 classes, and 20 examples are associated with that node, 12 of which belong to class1, 4 belong to class2 and 4 belong to class3; after the 3-way split the first node contains 8 examples 4 of which belong to class2… [↑](#footnote-ref-2)