Solution Sketches Midterm Exam B

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 [14]

Problem 4 --- Support Vector Machines [9]

Problem 5 --- EDA [12]

:

**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 information-gain[[1]](#footnote-1) for the following decision tree split for a dataset containing 3 classes (give the formula and compute the exact value; just giving the formula will only obtain partial credit) [4]

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

(1,1,0)

Entropy before: H(1/4.1/4.1/2)=3\*1/2=1.5

Entropy after 0+0.5\*H(1/2,1/2)=0.5

Entropy Gain= 1.5-0.5=1

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]

same

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

same

**2) Miscellaneous Questions**

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

The training set is used to learn the model.

validation sets are to select hyper parameters during training

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

0.95

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]

same

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]

same

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

Same

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

The following confusion Matrix of a classification model for distinguishing apples from oranges and mangos is given below:

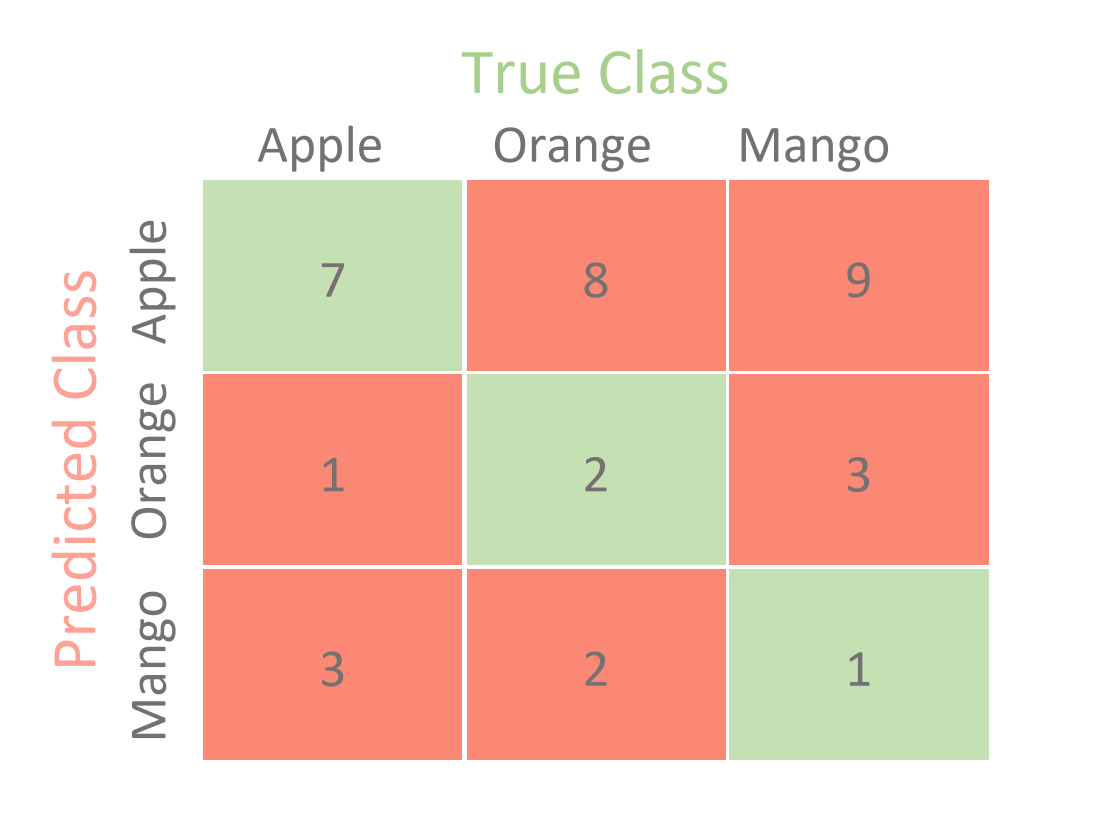
What is the accuracy of the classification model; what is its precision for class orange? What is its recall for class orange? It is okay to represent you answers as fractions; e.g. 17/36! [4; 1+1.5+1.5]

Accuracy: 10/36

Precision Orange: 2/6

Recall Orange: 2/12

No partial credit



**3) Neural Networks [12]**

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

wA,B=0.2

ΔB=0.5

A B

Assume that B is an intermediate node of a neural network, the forward propagation activation values of nodes aA and aB are 0.2 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 0.5, 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\*0.2\*0.5=0.2+0.05=0.25 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! [2]

same

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, taken from the figure below. [3]

wA,B=0.4

ΔB=0.5

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

Accidentally the same

Problem 3 continued

d) You train a neural network with a batch size of 200 (e.g. for a training set containing 1000 examples) What does this mean? [2]

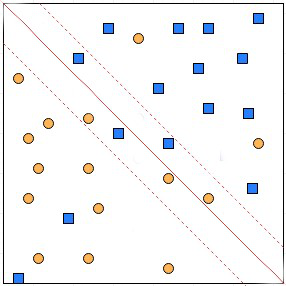
2000 samples (1/5 for the training set) will be passed through to the network at one time.

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

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

svn-equation

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]



same

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

same

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\*4 – A3\*5 + 11

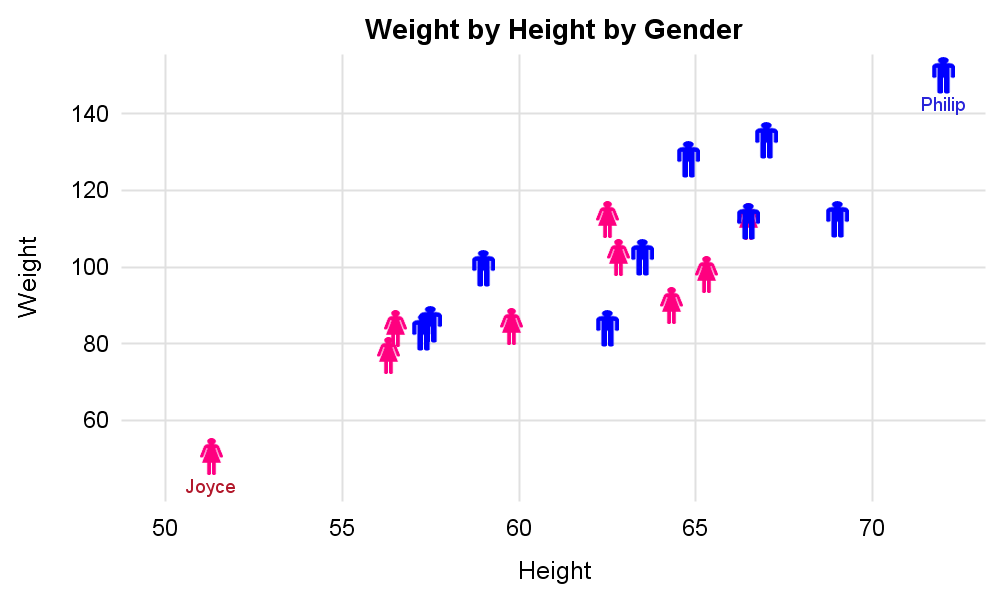
(0,1,3) that is, A1=0, A2=1 and A3=3.

Many other correct answers!

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

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]



Same

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

> x<-c (6,4,2,2,8,8,10,10,12,12,29,31)

> 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=9 [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: -6

Upper whisker: 12

Outliers: 29, 31

1. Entropy before (computed using the H-function) and after the split. There are 4 examples associated with the node to be split, one belong to class1 and class2 each, and two examples belong to class3. [↑](#footnote-ref-1)