Christoph F. Eick

**Review October 8, 2020 for COSC 6335 Midterm1 Exam**

**1) Decision Trees/Classification [16]**

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

(5,3,12) (4,0,2)

(1,3,0)

 (0,0,10)

**GINI-gain=GINIbefore-GINIafter=G(5/20,3/20,12/20) – 6/20\*G(2/3,0,1/3)-4/20\*G(1/4,3/4,0)-0**

**Partially correct : at most one point!**

1. If the GINI value is 0, what does this mean? [1]

**The objects which are associated with that node are pure—all objects belong to the same class.**

c) What are the characteristics of overfitting when learning decision trees? Assume you observe overfitting, what could be done to learn a “better” decision tree? [5]

**overfitting: training error low[0.5], testing error not optimal[0.5], models is too complex—the decision tree has to many nodes[1]**

**what to do to deal with it:**

1. **increase the degree of pruning in the decision tree learning algorithms to obtain smaller decision trees [2]**
2. **increase the number of training examples [1]**

**Other answers might exist which might deserve some credit!**

d) Most machine learning approaches use training sets, test sets and validation sets to derive models. Describe the role each of the three sets plays! [4]

**Training set: used to learn the model [1.5]**

**Test set: used to evaluate the model, particularly its accuracy [1.5]**

**Validation set: used to determine the “best” input parameter(s) for the algorithm which learns the model; e.g. parameters which control the degree of pruning of a decision tree learning algorithm or C in the case of the soft margin support vector machine. [2]**

e) Compute H(1/2, 0, 1/8,1/8,1/8, 1/8)=

½\*log\_2(2)+0+4\*1/8\*log\_2(8)=1/2+0.5\*3=0.5+1.5=2

2) APRIORI [7]

a) What is the APRIORI property? Where is it used in the APRIORI algorithm? [3]

Let *i* be an interestingness measure, and X and Y are sets:

**X⊆Y🡺*i*(X)≥*i*(Y) [1]**

1. **When creating k+1-itemsets candidates from frequent k-itemsets [1]**
2. **For subset pruning [1]**

b) Assume the APRIORI algorithm identified the following 7 4-item sets that satisfy a user given support threshold: **abcd, abce, abcf, acde, adef, bcde, and bcef;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

**Candidate 5-itemsets generated are: abcde, abcdf, abcef**

**abcde misses abde**

**abcdf misses bcdf**

**abcef misses acef**

**All three 5-itemsets are pruned!**

**One error: at most 2.5 points; 2 errors: at most 0.5 points**

3) Exploratory Data Analysis [10]

a) What is the role and purpose of exploratory data analysis in a data mining project? [4]

**a) collect background knowledge for the task at hand b) assess difficulty of the task to be solved c) provide knowledge to help in tool selection d) assess data quality / validate data e) helps in forming hypotheses f) find unexpected issues, errors and patterns in the data.**

b) Interpret the following 2 histograms and their relationships which describe the male and female age distribution in the US, based on Census Data. [6]



**Both histograms: curves are continuous with no gabs, and somewhat smooth[1], bimodal with 2 (1??; 0??) not well separated maxima at 5-19 and 35-44 [1.5], values significantly drop beyond age 55[1]🡪skewed distribution**

**Comparison: Curves are somewhat similar until age 55 [1] (although there are more males initially[0.5]); decline in the male curve is significantly steeper---women live longer[1].** Other observations might receive credit; points will be subtracted if you write things which do not make any sense or are false

**4.**Assume we have a support vector machine model for a dataset containing attributes A and B, whose hyperplane is defined as follows:



Which of the following 5 training examples is the closest to this hyperplan (3 points):

1. **(A=1,B=2)**
2. (A=9,B=9)
3. (A=-2,B=3)
4. (A=0, B=1)
5. (A=-1,B=1)

5. Which of the following statements about N-fold-cross validation is false (2 points):

a. N-fold cross validation estimates the error to classify examples which were not used in training

**b. *When n-fold cross validation is used the folds should be partically overlaping***

c. In n-fold cross validation each example occurs in N-1 of the N training sets

**d.** N-fold-cross validation is the most common approach to determine the training accuracy of a model.

e. When n-fold cross-validation is used the dataset example should be assigned to the n-folds randomly,

6. Which of the following practices would you recommend to battle overfitting (2 points):

**a. Increase the number of training examples**

b. Use a faster computer

c. Reduce the number of training examples

d. Increase model complexity

7. Which statement of using SVMs in conjuction with kernels is false? (3 points)

a. Kernels map the original dataset into a higher dimensional space and then find a hyperplane in the mapped space

**b. Kernels map the original dataset into a higher dimensional space and then find a hyperplane in the original space**

c. Using kernels allows to obtain non linear decision boundaries for a classification problem

d. The kernel trick allows to perform computations in the original space and enhances speed of SVM learning.

8. The soft margin support vector machine solves the following optimization problem:



Also take a look at Figure Y below.



Fig. Y. SVM Decision Boundaries for a Dataset containing Two Classes.

Which of the following statements is false (4 points)?

a. There are six example with positve ξi values in the above figure.

b. In the figure above, for all green examples which are below the lower dotted line their ξi values are 0.

c. In general, all examples for which ξi is less than half of the margin of the SVM will be classified correctly.

d. In general, all examples for which ξi is zero will be classified correctly

e. In general, all example with positive ξi  values will be misclassified.

10. Knn

a) How does kNN (k-nearest-neighbor) predict the class label of a new example? [2]

Find the nearest k neighbor of the example which needs to be classified; take a major vote based on the class labels of the k-nearest neighbors found.

b) Assume you want to use a nearest neighbor classifier for a particular classification task. How would you approach the problem to choosing the parameter k of a nearest neighbor classifier? [3]

Use N-fold cross validation to assess the testing accuracy of kNN classifier for different k values; choose the k for your final classifier for which the testing accuracy is the highest!

e) What can be said about the number of decision boundaries a kNN classifier uses? [2]

**Many[2]; as many as N-1 where N is the number of training examples[1 extra point].**

Comment: Interpreting scatter plots and comparing box plots and histograms will be part of Midterm1; however, as there we a lot of examples discussed in the EDA lecture, Dr. Eick decided not to include it in this review!

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