Christoph F. Eick

Midterm1 Exam

Solution Sketches

COSC 3337 *Data Science I*

October 3, 2023

Your Name:

Your student id:

Problem 1 --- Supervised Learning and Decision Trees [14]

Problem 2 --- Miscellaneous Questions [14]

Problem 3 --- Support Vector Machines [12]

Problem 4 --- EDA [16]

Problem 5 --- Non-Parametric Density Estimation [8]

64:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 14% towards the course grade. The use of computers and cell phones is strictly prohibited.

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

1. **Decision Trees and Supervised Learning [14]**
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)! [6]

(5,3,2) (2,3,0)

(0,0,2)

(3,0,0)

Formula 3 points

G(5/10, 3/10, 2/10) – 5/10\*G(2/5, 3/5,0) – 2/10\*G(0/2, 0/2,2/2) – 3/10\*G(3/3, 0/3, 0/3)

~ G(0.5,0.3,0.2)-[0.5\*G(0.4,0.6,0) + 0.2\*G(0,0,1) + 0.3\*G(1,0,0)]

(1-0.5\*\*2-0.3\*\*2-0.2\*\*2)- 0.5\*(1-0.4\*\*2-0.6\*\*2)-0-0=

(1-0.25-0.09-0.04) – 0.5\* (1-0.16-0.36)=

0.62-0.5\*(0.48)= 0.62-0.24=19/50=0.38

One error at most 4 points

Two errors at most 1.5 points

3 errors no points

1. Assume you learn a decision tree for a dataset and you observe overfitting. What does this mean? What can be done to alleviate overfitting? [4]

The training error is low[1] or insufficient training examples[1] and the testing/generalization error is not optimal [1].

Increase size of training set [1]; reduce the number of nodes of the tree [1]

1. Compute H(1/8,1/16,1/16,0,1/2,0,1/4)! [3]

1/8\*log\_2(8) + 2\*1/16\*log\_2(16)+0+1/2\*log\_2(2) + ¼\*log\_2(4)

3/8 + ½ + ½ +1/2=15/8=1.875

If error(s) 0-1 points!

1. Assume you have a classification problem with 32 classes; what is the highest value the entropy function H can take? [1]

5

**2) Miscellaneous Questions [14]**

a) Given a training set T and distance function d for T; how does a k-nearest classifier determine the class of an example e? [2]

Computes the k-nearest examples in T to e and then conducts a majority vote based on the class of the k-nearest-neighbors to determine the class to be returned. [2]

Or calculates the Euclidean distance and uses k “closest” points for classification by taking majority vote

1. What happens if you use a k-nearest neighbor classifier with large k values; e.g, for a dataset with 200 example you use k=185. What happens in this case? In general, is k=185 a good choice for k? [2]

The model will underfit and all new examples will likely receive the same class[1] or Determines the majority class of the dataset [1]

No [1] K is too much above the mean number of examples in the dataset.

Other answers might deserve partial credit.

Problem 2 continued

1. Assume we have a dataset with an attribute A with a mean value 8(μ=8) and standard deviation 4(σ=4). According to the 68–95–99.7 rule, what is the probability that a value of attribute A is between 0 and 16? [2]

0.95 No partial credit for other answer.

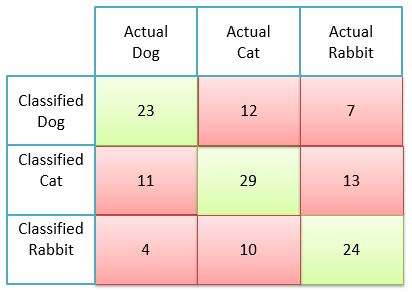
1. Assume you have a dataset with numerical attributes A and B which have a correlation of -0.098; what does this say about the relationship of attributes A and B? [2]

There’s little to no linear relationship between both Attributes [2]

OR There is only a weak linear relationship between attributes A and B. When A goes up B goes down and vice versa [2]

e) A confusion Matrix of a classification model for distinguishing dogs, cats and rabbits is given below:

What is the accuracy of the classification model; what is its precision for class rabbit? What is its recall for class rabbit? It is okay to represent your answers as fractions; e.g. 17/36! [3]



Accuracy =(23+29+24)/ (23+12+7+11+29+13+4+10+24)=76/133=4/7=0.5714 [1]

Precision Rabbit= 24/(4+10+24)=24/38=12/19=0.6316 [1]

Recall Rabbit = 24/(7+13+24)=24/44=6/11=0.5454 [1]

f) Most publications in medicine report recall and precision and not accuracy for the experiments they conducted. Why do you believe this is the case? [3]

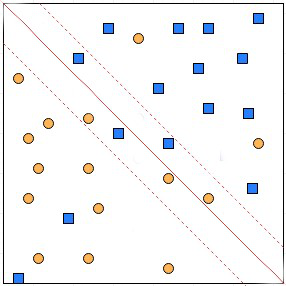
False positive and false negative have often different cost associated with the particular wrong decision, making it necessary to distinguish errors with respect to false negatives and false positive.[3] **Other answer might deserve partial credit.**

**3) Support Vector Machines [12]**

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

svn-equation

What is the purpose of the first term in the above formula? [1] What does ξi measure?[2] How many examples in the figure are misclassified?[1] Assume example j has a positive ξj value; does this mean that example j will be misclassified; give a reason for your answer! [3]



It measures the width of the inverse margin [1]; if they say margin only 0.5 points);

the error of the j’s example which is 0 if the example is on the correct side of the example’s class hyperplane and the distance to the example’s class hyperplane[2]; if they correctly visualize the error in the above figure correctly you can also give them 2 points.

6 examples are misclassified[1]

No; example e can be between the hyperplane and the hyperplane of e’s class in which case the have a positive error but e will be classified correctly [3]

Or

-The purpose of the first term is to minimize the inverse margin [1]

- error term measure the distance of the data points to the hyperplane [2]

- 6 examples are misclassified [1]

- No, positive error values does not mean it is misclassified [3]

b) SVMs have been successfully used in conjunction with Kernels. Why do you believe this is the case? [3]

Mapping the dataset to a higher dimensional space [1] creates more ways to separate the example leading to a higher SVM accuracy or little to no error[2].

Other answers might deserve partial or full credit!

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\*5 + A2\*3 – A3\*2 + 1

How is this hyperplane used to classify an example; e.g. (A1=1,A2=0,A3=6)? [2]

You plug in the attribute values into the equation and determine the class based on the sign of the number you get.[2]

Alternatively, they can say the computation leads to 5-12+1=-6 and the negative class is selected.

4) Data Analysis [16]

a) Interpret the supervised scatter plot, depicting instances of classes

called “Group 1’, ‘Group 2’ and ‘Group 3’ with their respective values of attributes named x and y. Characterize the distribution of the instances of each class in the attribute space. Assess the difficulty of the classification problem of predicting the 3 classes using the attributes x and y! [10]

A group of dots in a row

Description automatically generated with medium confidence

Group 1 is concentrated in the [-0.2, 0.3 or -0.2,0.6] x [-4, 5] rectangle, Group 2 is concentrated in the [0.7, 1.2] x [0, 9] Rectangle, Group 3 is concentrated in the [0.3, 1.0] x [-2, 5 or -6,7] Rectange [3];

all three classes have a unimodal distribution [1.5].

It is trivial to separate Group 1 and Group 3[1];

Group 1 and Group 3 are well separated except in the [0.2, 0.5] x [-5, 5] region[1];

Group 2 and 3 are also separable but with more effort: there is more overlap between Group 2 and Group 3, as several Group 2 examples occupy in the main Group 3 region, but the two classes are still separable with some effort[1].

Attribute x is quite useful for the classification problem is using x=0.4 and x=0.8 accomplishes good separation between the 3 classes or Attribute x is quite useful for the classification problem to separate the 3 classes, x < 0.5 – Group 1, 0.5 < x < 1 – Group 3, 0.8 < x < 1.2 – Group 2 [3];

attribute y is less useful except the y values of Group 1 seem to be lower than the y values of Group 2 which are usually lower than the y values of group 3[1].

Overall, the classification problem has moderate difficulty; however, a few challenges exist in the who overlap regions, identified earlier [1.5].

At most 10 points; other observations might deserve some credit. Associate negative points with false observations!

b) Boxplots [6]

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

> x<-c (15,4,3,3,9,10,12,12,12,12,19,29)

> boxplot(x)

What is the median for the attribute x?[1] What is the IQR for the attribute x?[1.5] What is the position of the upper whisker of the box plot created for attribute x?[1.5] Are there any outliers in the dataset[2] Assume that outliers are values which are 1.5 IQR above the upper box boundary or 1.5 IQR below the lower box boundary. Justify your answers for the last two questions!

Median: 11 [1]

IQR: 12-3=9 [1.5]

Upper Whisker: 19 [1.5]

Outliers; -15 and 29 [2]

The upper whisker is 19 as it is highest non-outlier. With the upper boundary being 25.5 and the lower being -10.5, the two outliers are -15 and 29.

Side computation: Range of non-outliers for x: [-10.5,25.5] ~ (3-13.5=-10.5 12+13.5=25.5)

No partial credit; except give a single point if they identify one of the two outliers.

5) Non-parametric Density Estimation 8]

a) Describe in a single natural language sentence without formulas how nonparametric density estimation computes the density for a query point x! [3]

Density is computed by adding the influences of each point in the dataset with respect to the query point x [3]

Other answers might deserve partial credit.

b) Assume a dataset O={x1,x2,x3,x4} with data points x1=(1,1), x2=(4,4), x3=(6,4), x4=(4,6) is given; moreover, assume Manhattan distance[[3]](#footnote-3) is used as the distance function and q1=(5,5) is a query point. Compute fGauss (q1) assuming bandwidth σ=1! [5]

Remark: it is okay to use an expanded formula as your answer; e.g “e-12 + e-2.5…” as your answer; it is not necessary to report the exact value!

d(x1, q1) = |5-1| + |5-1|=8, d(x2, q1) = 2, d(x3, q1) = 2, d(x4, q1) = 2

A green background with white text

Description automatically generated

Gaussss

fGauss ((5,5)) = e-64/2 + e-4/2  + e-4/2 + e-4/2 = e-32 + 3\* e-2 ~ 0.406 [5]

Can give up to 3 points partial credit, but 3 points only if minor errors.

1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)
2. There are 3 classes, and 10 examples are associated with that node, 5 of which belong to class1, 3 belong to class2 and 2 belong to class3; after the 3-way split the first node contains 5 examples 2 of which belong to class1 and 3 of which belong to class2. [↑](#footnote-ref-2)
3. d((x1,y1),(x2,y2))= |x1-x2| + |y1-y2| [↑](#footnote-ref-3)