Solution Sketches Midterm Exam1

COSC 3337 *Data Science I*

October 2, 2019

Your Name:

Your student id:

Problem 1 --- Supervised Learning and Decision Trees [10]

Problem 2 --- Basic Statistics [10]

Problem 3 --- Data Visualization [11]

Problem 4 --- Non-Parametric Density Estimation [10]

Problem 5 --- Data Analysis [19]

:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 15% towards the course grade. The use of computers and cell phones is strictly prohibited. The exam is slightly to long

1. **Decision Trees and Supervised Learning [10]**
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)! [7]

(12,6,6) (6,6,0)

( 6,0,0)

(0,0,6)

Gini-Gain

G(0.5,0.25,0.25)- (0.5\*G(0.5,0.5,0)+0.25\*0+0.25\*0)=(1-6/16)-0.5\*0.5=10/16-1/4=6/16=3/8

Wrong Formula: at most 3 points; correct formula but wrong value 4.5 points.

1. An unnamed scientist uses training and testsets for supervised learning that share 25% of examples—25% of the examples in the training set belong to the corresponding testset. Explain, why this is not a good practice! [3]

*The purpose of having a test set is to assess the accuracy of the learnt model on unseen examples that have not been used in training. Having training examples in the testset defeats the purpose of “a testset” as accuracy is no longer assessed solely using examples that have not been used in training,*

**2) Basic Statistics [10]**

a) In Data Science raw data sets are frequently z-scored before applying a particular analysis technique to them; what is the motivation for doing that? [2]

*To normalize and make attributes equally important or by alleviating the fact that different attributes have a different scale,.*

1. The correlation between attribute A and attribute B is -0.97; what does this tell you about the relationship of the two attributes? [2]

*A and B have a strong linear relationship[1]; if the value of attribute A is high the value of attribute B is low and vice versa[1].*

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 4 and 12? [3]

P(4≤A≤12)=0.95 No partial credit!

1. Does the 68–95–99.7 rule always compute the correct probability in the above example? Give a reason for your answer! [3]

*No [1]; the formula assumes that attributes A follows a Normal distribution N(μ,σ); if A’s probability distribution is not a Normal the suggested probabilities are often incorrect[2]!*

**3) Data Visualization [11+ up to 3 extra points]**

1. What are the different types of goals of data visualization? [3]

* *to explore •Nothing is known, •Vis. used for data exploration –…*
* *to analyze •There are hypotheses, •Vis. used for Verification or Falsification –…*
* *to present •“everything” known about the data, •Vis. used for Communication of Results*

b) Explain what the following C.R.A.P. Design Principles emphasize[4]:

* Contrast [2]
* Repitition [2]

*Contrast is all about making distinct elements stand out. Contrast is used to drive a user’s attention to specific elements in a design.[2]* Other answers might deserve partial or full credit.

*Repetition is how you maintain consistency in a design. It helps users familiarize with the way information is presented to them.[2]* Other answers might deserve partial or full credit.

1. What is the aspect ratio? One challenge when visualizing line charts is choosing the proper aspect ratio. What things should be considered when choosing the proper aspect ratio for a line chart? [4+up to 3 extrapoints for interesting “other” considerations]

The ratio between the width and the height of a rectangle is called its aspect ratio[1].

*Considerations:*

*Bank at 45 degrees [2]*

*Considering constraints of the available space/required formats to fit the display(e.g width of a given column in an IEEE ) / [1]*

*Data Ink Ratio[1]*

… other considerations might deserve credit!

**4) Non-Parametric Density Estimation [10]**

1. Assume we use 2D non-parametric density estimation for a dataset D={(0,0), (2,2), (0,2)} and h=1. Give the formula that computes the density in the query points (3,3)---give the formula of influences of each point[[3]](#footnote-3)! [6]

ϕO(v) (General Formula)

ϕD((3,3) (e-9+e-1+e-5)

One error: at most 4 points

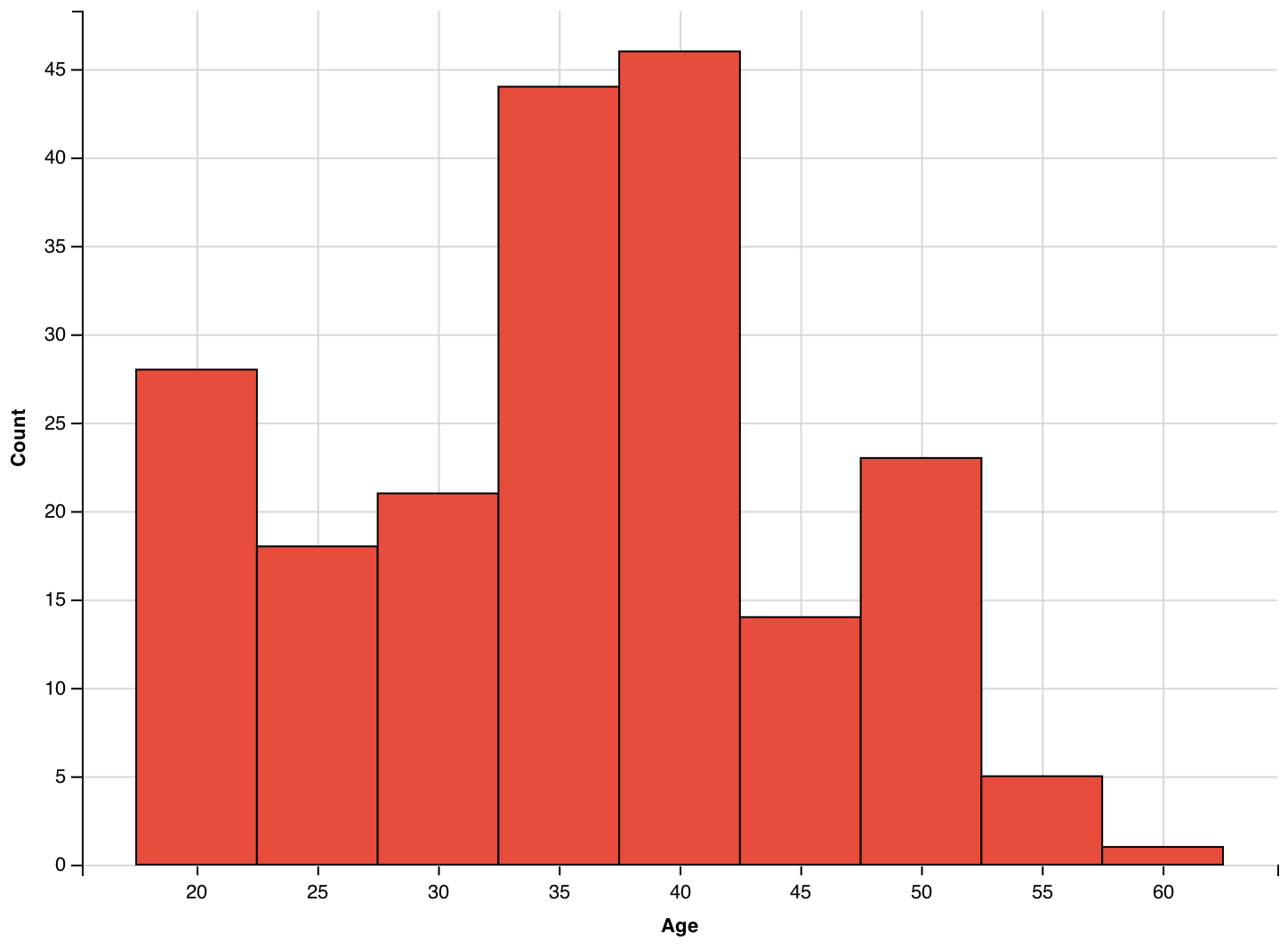
More than 1 error: 0 or 1 point.

1. What role does the bandwidth h (sometimes named σ) play in non-parametric density estimation? What are differences between density functions that have been constructed using high bandwidth values and density functions that have been constructed using low bandwidth value? [4]

*The bandwidth determines how quickly the influence of a point (in the dataset) on the query point decreases with the increase of distance [1]. h high few hills smooth display [1.5] h low rugged display with a lot variation/many hills [1.5] other correct answer: h high captures regional variation, h low captures local variation.*

5) Data Analysis [19]

a) Interpret the following histogram of the age distribution of a group of people! [5]



3-modal [1.5] with peaks at [17.5-22.5],[32.5,42.5], [47.5, 52.5] [1.5]

The [32.5, 42.5] peak is significantly higher than the other two peaks [1]

Not many people in the group with age 55+ [1]

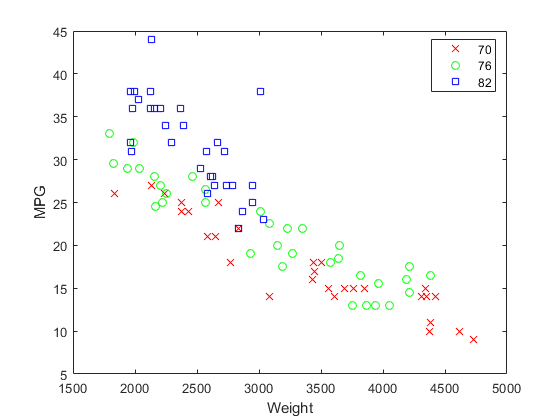
Assessing the Symmetry/skewedness of the histogram [1]

No gaps [1]

At most 5 points: Other observation might deserve credit!

Problem 5 continued

1. Interpret the supervised scatter plot depicted below that consists of instances of 3 classes called 70, 76 and 82; moreover, assess the difficulty of separating instances of the 3 classes using attributes MPG and Weight based on the scatter plot! [8]



Some observations:

Class blue unimodal (bimodal is also okay) [1]

Class green unimodal (or bi-modal with gap near (25, 2800) or trimodal//5 modal[[4]](#footnote-4) ) [1]

Class red multi-modal [1]

Class blue is easy to separate from class red[1.5], and can be separated with few errors from class green[1.5]

Instances of class red tends to have lower MGP values than class green, but separation the two classes is quite challenging.[2]

Saying where the instances of the 3 classes are concentrated might deserve up to 1.5 points.

Other observations might deserve credit.

At most 8 points!

Problem 5 continued

c) The boxplot depicted below has been created using the following R-code for an attribute x:

> x<-c (1,2,2,2,8,8,8,10,14,14,18,37)

> boxplot(x)

What is the median for the attribute x? What is the IQR for the attribute x? The higher whisker of the boxplot as at 18; what does this tell you? According the boxplot 18 is not an outlier and 37 as an outlier; why do you believe this is the case? [6]

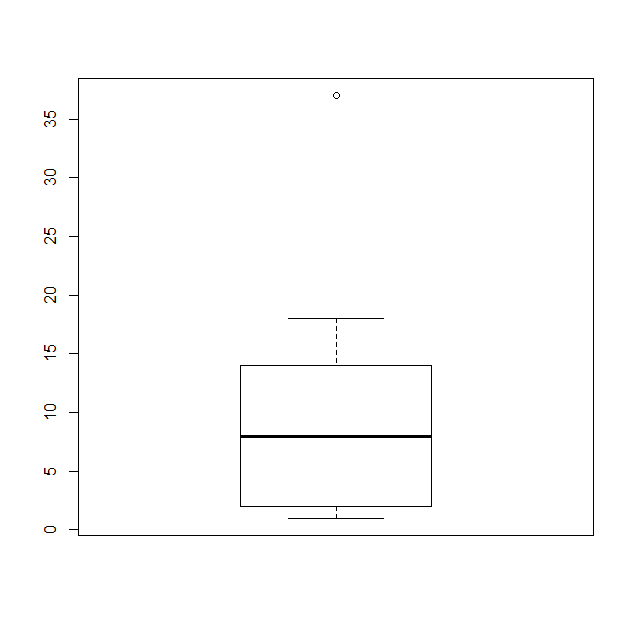
Median=8 [1]

IQR=14-2=12 [1]

18 is the highest non outlier value for attribute x [1.5]

High outliers are 1.5 IQR above the 25% percentile; in our case 14+1.5\*12=32; that is, all points that are above 32 will be depicted as outliers in the plot [2.5]

No partial credit for the first 3 questions!!



1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)
2. There are 3 classes, and 24 examples are associated with that node, 12 of which belong to class1, 6 belong to class2 and 6 belong; after the 3-way split the first node contains 12 examples 6 of which belong to class1… [↑](#footnote-ref-2)
3. But you do not need to compute their exact values! [↑](#footnote-ref-3)
4. Hard to say---all answers “carry some correctness”! [↑](#footnote-ref-4)