Christoph F. Eick

**Review Sept. 30 2025 for the COSC 3337**

**Exam1 on Oct. 2, 2025**

**Other important tasks not included here for Midterm1 include: Interpreting boxplots; comparing histograms and boxplots,…**

**1) Neural Networks**

a) 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, for those which are known! [4]

wA,B=0.5

ΔB=0.2

A B

wC**,A**=1 wD,A=0.2

C D

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.2\*0.4=g’(0.24)\*0.5\*0.2

If not correct at most 1 point partial credit

b) The step width in NN gradient descent search varies, depending on various factors including depending on the derivative of the activation function and the activation of the input node. Explain why each of those two variations is desirable? [4]

If the gradient of the activation function is low small steps are used to not overshoot the local minimum which is characterized by a gradient of zero, [2]

If the activation of the input is low the link has very little influence on the activation of the output node; consequently, the weight of the connection is only changed a little or not at all. [2]

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

1. Compute the Information gain for the following decision tree split (just giving the formula and compute its value 6]:

(4,4,4,4) (4,4,0.0)

(0,0,4,0)

(0,0,0.4)

**Information Gain= H\_Before-H\_After= H(1/4.1/4.1/4,1/4)-1/2\*H(0.5,0.5,0,0)-1/4\*0-1/4\*0=4\*1/2- ½\*(1/2+1/2)=1.5**

**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

f) Why is pruning important when using decision trees? What is the difference between pre-pruning and post pruning? [3]

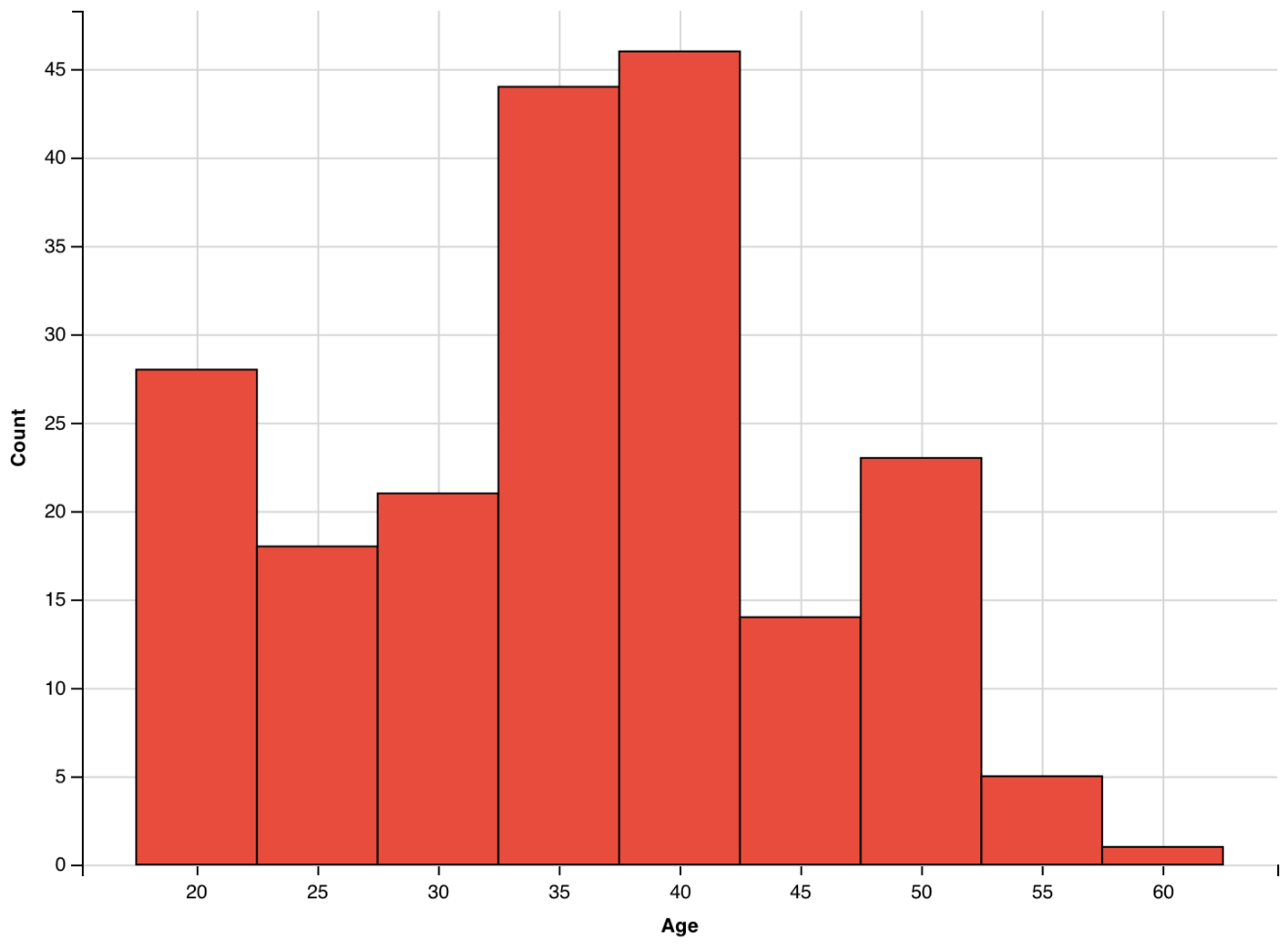
To avoid overfitting[1] Pre-pruning uses “stronger” termination conditions for the decision tree induction algorithm to obtain smaller trees[1] Post-pruning grows a large decision tree and then reduces its size (usually relying on a validation set)

g) Assume you apply the decision tree induction algorithms without any pre-pruning to a dataset which does not contain inconsistent[[1]](#footnote-1) examples. What can be said about the training error of the learnt decision tree? Give a reason for your answer.

100% training accuracy. Reason: Assume the accuracy is not 100%; then the learnt decision tree must contain a leaf node with associated training examples e1 and e2 having different class labels; however, as the training set does not contain any inconsistent examples, e1 and e2 have to disagree in at least one attribute X; however, this is impossible as the decision tree induction algorithm would have separated e1 and e2 by adding another test which separates e1 and e2 based on attribute X. Consequently, the training accuracy is 100%.

**3. Histograms**

Interpret the following histogram!



The histogram is tri-modal; it has 3 peaks in the age ranges 17.5-22.5 32.5-42.5 and 47.5-52.5. The histogram is not skewed; the histogram shapes left and right of the 32.5-42.5 peak are somewhat similar. There are no gaps! Ages in the range 57.5-62.5 are outliers.

Other things to look at: particularly in histograms with more bins🡪look at differences in slop going down a peak

approach has been a “success story”, frequently led to obtaining high accuracy classifier? [3]

The dataset is mapped to a higher dimensional space [1] and a hyperplane is found in the mapped space not the original space [1]

As the mapped space is higher dimensional there are more ways / more possible hyperplanes to separate the example of the two classes with no error or a low error. [3]

As kernel functions are usually non-linear, decision boundaries in the mapped space correspond to non-straight line decision boundaries in the original space, facilitating finding good hyperplanes [2]

Other answers to the second question might deserve partial credit!

**4) 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. 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 4+2σ=8/ 12-2σ=8

d) 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]!*

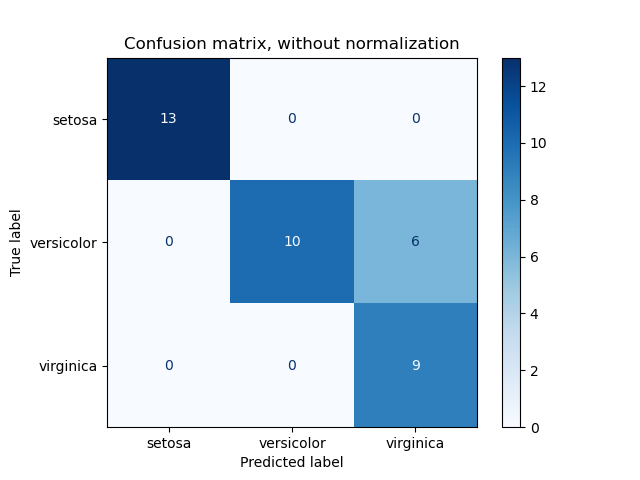
*e)* In Data Science raw data sets are frequently z-scored before applying a particular analysis technique to them; what is the motivation for doing that? [2]

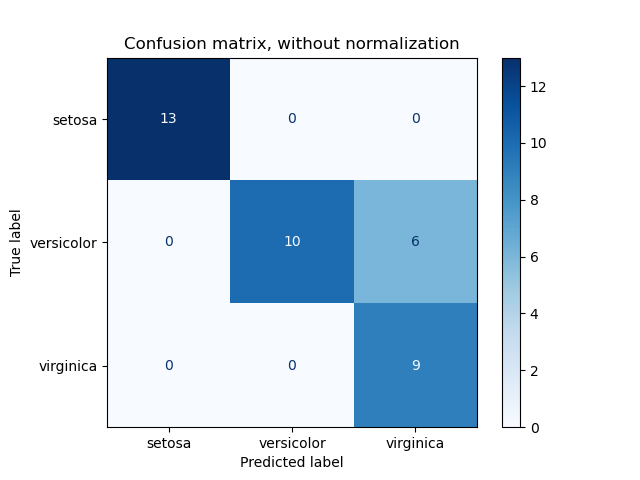
*To make different attributes equally important to alleviate the fact that different attributes usually differ in their scale.*

5. Classification Model Performance Evaluation Measures

The following confusion Matrix of a classification model of the IRIS flower dataset is given below:

What is the accuracy of the classification model; what is its precision for class versicolor? What is its recall for class versicolor?





Accuracy: 13+10+9/13+10+9+6=32/38=82.4%

Precision Versicolor (if the classifier predicts Versicolor; how often is this decision correct?)= 10/10=100%

Recall Versicolor (if the actual class is Versicolor; how often makes the classifier the correct decision?): 10/16=62.5%

For Virginica, on the other hand, the recall is 100% but the precision is quite low.

For Setosas the recall and precision are both 100%.

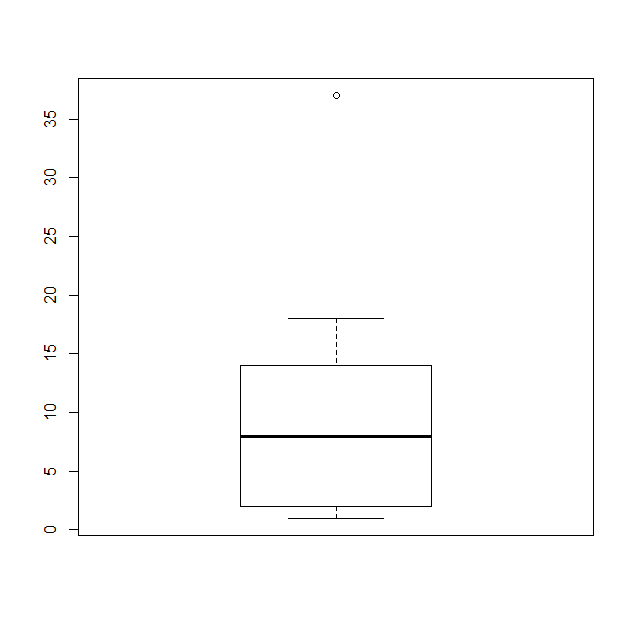
6. Boxplots

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 75% 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]

7. K-Nearest Neighbor

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 in the training set; take a majority 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!

c) 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 maybe comparing box plots and histograms will be part of the midterm exam; however, as there we a lot of examples discussed in the EDA lecture, Dr. Eick decided not to include these topics in this review!

1. 2 examples are inconsistent if they disagree in their class attribute but agree in all their other attributes; e.g. (0.4,0.6,0.7,0.6, Setosa) and (0.4,0.6,0.7,0.6,Versicolor) in an Iris flower dataset. [↑](#footnote-ref-1)