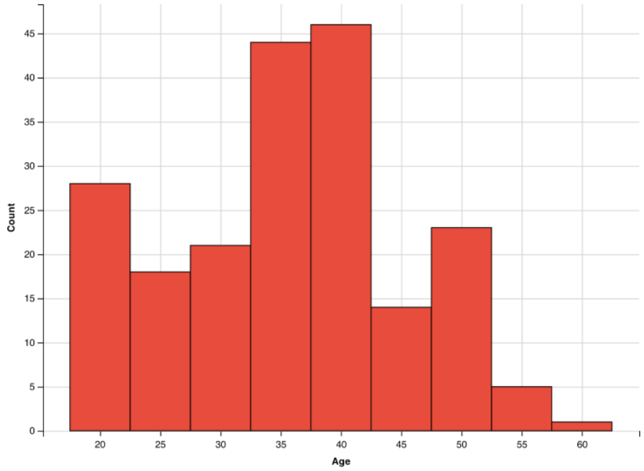
October 3, 2022
Data Science
Christopher Eick

Past Examinations and Review Content for Midterm 1 (COSC 3337)

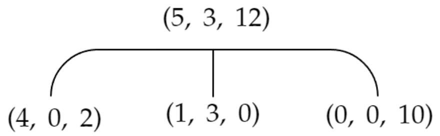
1. Histograms:
   Directions: *Interpret the following histogram below:*



   We observe that the histogram is tri-modal (meaning is has 3 distinguishable peaks) observed in the bins $[17.25, 22.5]$, $[32.5, 42.5]$, 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 $[57.5 - 62.5]$ appear to be outliers.

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

2. Decision Trees and Classification:
   a. Directions: *Compute the GINI-gain[1] for the following decision tree split (just giving the formula is fine!):*



   Suppose that we denote GINI gain as γ, the GINI coeffecient before the split as α_{before}, the GINI coeffecient after the split as α_{after}, and the GINI function of a bin as $G(x, y, z)$ where $x, y, z \in \mathbb{N}$ represent the number of data objects. Then we have that:

$$\gamma = \alpha_{before} - \alpha_{after} = G\left(\frac{5}{20}, \frac{5}{3}, \frac{5}{12}\right) - \left[\frac{6}{20} \cdot G\left(\frac{4}{6}, 0, \frac{2}{6}\right) + \frac{4}{20} \cdot G\left(\frac{1}{4}, \frac{3}{4}0\right) + 0\right]$$

   Note that $G(0, 0, 10) = 0$ cause all the values belong to one class (class 3).

   b. Directions: *If the GINI value is 0, what does this mean?*

[1] (GINI before the split) minus (GINI after the split)

It means that all the data objects in the following bin belong to one class. That is, it is homogenous.

c. Directions: *What are the characteristics of overfitting when learning decision trees? Assume you observe overfitting, what could be done to learn a "better" decision tree*

Some characteristics of an overfitted tree are: the training error low, the testing error not optimal, the models is too complex—the decision tree has to many nodes.

We can deal with a overtrained tree in two ways:

1. Increase the degree of pruning in the decision tree learning algorithms to obtain smaller decision trees.
2. Increase the number of training examples.

Remark: Other answers might exist which might deserve some credit!

d. Directions: *Most machine learning approaches use training sets, test sets and validation sets to derive models. Describe the role each of the three sets plays!*

Training set: numerical sets used to learn/derive the model.
Testing set: numerical sets used to evaluate the model, particularly its accuracy.
Validation set: numerical sets 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.

e. Directions: *Compute $H\left(\frac{1}{2},0,\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8}\right)$, where $H(x)$ is the entropy function of and $x \in \mathbb{R}^n$ is a vector, consisting of n values.*

Recall that the entropy formula is either given by $e(x) = -\sum_{i=1}^{n} xlog_2 x$ or $e(x) = \sum_{i=1}^{n} xlog_2\left(\frac{1}{x}\right)$. By using the latter formula, we have that $e(x) = \frac{1}{2}\cdot \log_2(2) + 0 + 4\cdot\left[\frac{1}{8}\cdot \log_2(8)\right] = \frac{1}{2} + 3\cdot\frac{1}{2} = 2$.

f. Directions: *Why is pruning important when using decision trees? What is the difference between pre-pruning and post pruning?*
To avoid overfitting, pre-pruning uses "stronger" termination conditions for the decision tree induction algorithm to obtain smaller trees and post-pruning grows a large decision tree and then reduces its size (usually relying on a validation set).

3. Support Vector Machines

a. Directions: *Assume we have a support vector machine model for a dataset containing attributes x and y, whose hyperplane is defined as follows:*

$$3x - 2y + 1 = 0$$

Which of the following 5 training examples is the closest to this hyperplane?:

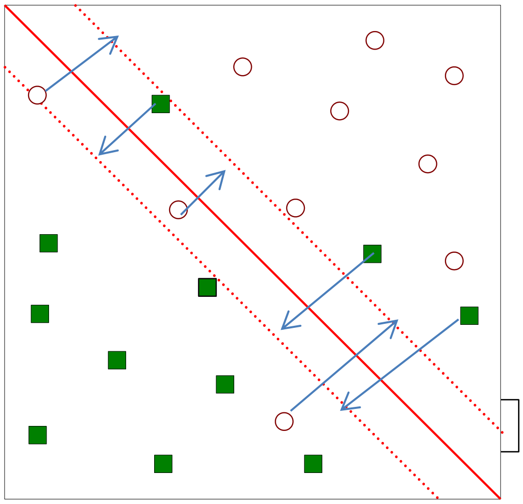
a. $(\boldsymbol{x = 1, y = 2})$
b. $(x = 9, y = 9)$
c. $(x = -2, y = 3)$
d. $(x = 0, y = 1)$
e. $(x = -1, y = 1)$

b. Directions: *The soft margin support vector machine solves the following optimization problem:*

$$argmin\left(\frac{1}{2}||w||^2\right) + C\sum_i \xi_i$$

Where $c_i(w \cdot x - b) \geq 1 - \xi_i$ and $1 \leq i \leq n$. Below, the ξ_i values (which are the length of the depicted arrows) are illustrated:

All other points have ξ_i values of 0!



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

a. There are six example with positive ξ_i values in the above figure.

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

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

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

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

c. Directions: *Support vector machines are frequently successfully used in conjunction with kernels—how does this approach work? Why do you believe this approach has been a "success story", frequently led to obtaining high accuracy classifier?*

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

   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.

   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.

Other answers to the second question might deserve partial credit!

4. Basic Statistics

a. Directions: *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?*

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

b. Directions: *The correlation between attribute A and attribute B is -0.97; what does this tell you about the relationship of the two attributes?*

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

c. Directions: *Assume we have a dataset with an attribute A with a mean value 8 ($\mu = 8$) and standard deviation 2 ($\sigma = 2$). According to the 68–95–99.7 rule, what is the probability that a value of attribute A is between 4 and 12?*

   Observe the following:

$$\mu + 2\sigma = 8 + 2 \cdot 2 = 12$$

and

$$\mu - 2\sigma = 8 - 2 \cdot 2 = 4$$

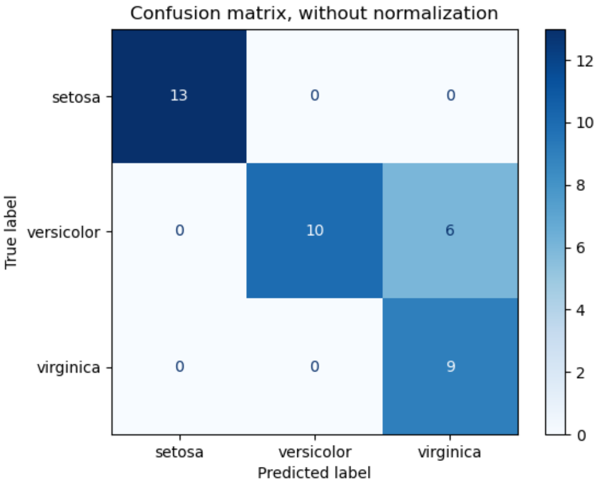
   By the $68\% - 95\% - 99.7\%$ rule for normal distributions, we have:

d. Directions: *Does the 68–95–99.7 rule always compute the correct probability in the above example? Give a reason for your answer!*

   No! The formula assumes that attributes A follows a normal distribution. However, if A's probability distribution is not a normal the suggested probabilities are often incorrect (e.g.) could be a Maxwell-Boltzmann distribution or Chi Distribution.

5. Classification Model Performance Evaluation Measures

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



Past Examinations and Review Content for Midterm 1 (COSC 3337)

Directions: *Determine the following:*

(a) What is the accuracy of the classification model;
(b) What is its precision for class versicolor?
(c) What is its recall for class versicolor?

We have that:

(a) The accuracy of the confusion matrix is given by: $\frac{13+10+9}{(13+10+9)+6} = \frac{32}{38} \approx 0.824$. Thus we have an accuracy of 82.4%.
(b) We have that the precision of the versicolor class (if the classifier predicts versicolor; how often is this decision correct?) is given by $\frac{10}{10} = 1$. Thus we have a 100% precision.
(c) We have that the recall of the versicolor class (if the actual class is Versicolor; how often makes the classifier the correct decision?) is given by $\frac{10}{16} = 0.625$. Thus, we have a 62.5% recall.

Remark(s):

- 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

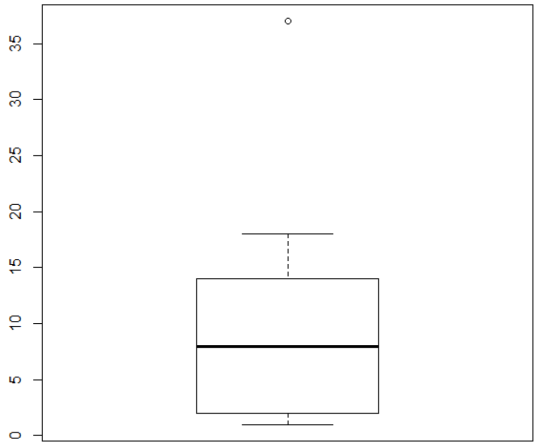
The boxplot depicted below has been created using the following R-code for an attribute x:

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

Directions: *Answer the following questions:*

(a) *What is the median for the attribute x?*
(b) *What is the IQR for the attribute x?*
(c) *The higher whisker of the boxplot as at* 18*; what does this tell you? According to the boxplot,* 18 *is not an outlier but* 37 *is an outlier; why do you believe this is the case?*

Past Examinations and Review Content for Midterm 1 (COSC 3337)



We note the following:

(a) We observe that the median is 8 (we can't exactly tell from the diagram, but by observing the R code, we deduce it is 8).

(b) We observe that the interquartile range is given by the value at $75\% - 25\%$. Fortunately, in our R code, the values are already sorted. We have 12 values overall. We note that $0.25 \cdot 12 = \frac{1}{4} \cdot 12 = 4$ and $0.75 \cdot 12 = \frac{3}{4} \cdot 12 = 8$. So we use the values at the 4 $-$th position and the 8 $-$th position. That is 2 and 14. Thus, the interquartile range is given by $14 - 2 = 12$.

(c) 18 is the highest non outlier value for attribute x. High outliers are $1.5 \cdot \text{IQR}$ above the 75% percentile; in our case $14 + 1.5 \cdot 12 = 32$; that is, all points that are above 32 will be depicted as outliers in the plot.

7. $k -$ nearest neighbors

a. Directions: *How does* kNN *(*k *- nearest-neighbor) predict the class label of a new example?*

It finds 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. Directions: *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?*

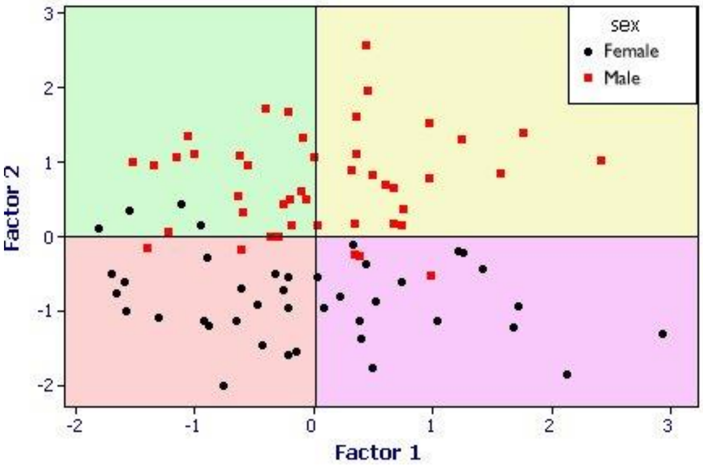
Use n −fold cross validation to assess the testing accuracy of kNN classifier for different k values; choose the k for your final classifier for which the testing accuracy is the highest!

c. Directions: *What can be said about the number of decision boundaries a k nearest neighbors classifier uses?*

Many; as many as $n - 1$ where n is the number of training examples.

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!

d. Directions: Interpret the supervised scatter plot depicted below; moreover, assess the difficulty of separating males from females using Factor 1 and Factor 2 based on the scatter plot!



Factor2 does mostly a good job in separating females and males; there is only overlap close to 0 [3]; Factor1 does a poor job separating the 2 classes/

e. What is the goal of density estimation? [2]

To construct an estimate based on observed data of an unobservable of an unobservable underlying probability density function. That is, the goal is to derive a density functions that best approximates the distribution.

8. Similarity Assessment

Past Examinations and Review Content for Midterm 1 (COSC 3337)

Directions: *Design a distance function to assess the similarity of electricity company customers; each customer is characterized by the following attributes:*

a. Social Security Number (SSN)
b. On Time Payment History (OPH) which is ordinal attribute with values ‘very good’, ‘good, ‘medium’, ‘poor’ and ‘very poor’.
c. Power-used (which is a real number with mean 100, standard deviation is 10, and the maximum 1000 and minimum 20)
d. Gender as a nominal attribute taking values in $\{male, female\}$.

Assume that the attributes OPH and Power – Used are of major importance and the attribute Gender is of a minor importance when assessing the similarity between customers. Using your distance function compute the distance between the following 2 customers: $c_1 = (111111111,\ ‘good’, 150, male)$ *and* $c_2 = (222222222, ‘very\ poor’, 50, female)$*!*

One possible answer: Ignore SSN as it is not important.

First, we normalize \`amount_spent\` using z - score and find distance by L_1 norm. We convert the OPH rating values: ‘very good’, ‘good, ‘medium’, ‘poor’, and ‘very poor’ to a 4:0 scale using a function, say $\phi(x)$. Then, we find distance by taking L_1 norm and dividing by range. Then we assign weights 0.4 to OPH, 0.4 to power-used and 0.2 to gender

We define the following functions:

$$d_{gender}(a,b) = 1 - \delta_{ab}$$

Where $1 - \delta_{ab}$ is the Kronecker-delta function (so if $a = b,$ then $\delta = 1$). Thus, we have:

$$d(u,v) = 0.4 \cdot \left|\frac{u_{powerUsed} - 100}{10} - \frac{v_{powerUsed} - 100}{10}\right| + 0.4 \cdot \frac{|\phi(u_{OPH}) - \phi(v_{OPH})|}{4} + 0.2 \cdot d_{gender}\big(u_{gender}, v_{gender}\big)$$

So, in our example, we would have:

$$d(c_1, c_2) = 0.4\left|\frac{150 - 100}{10} - \frac{50 - 100}{10}\right| 0.4 \cdot \left|\frac{3 - 0}{4}\right| + 0.2 \cdot 1 = 0.3 + 4 + 0.2 = 4.5$$

9. Data Mining in General

Past Examinations and Review Content for Midterm 1 (COSC 3337)

Directions: *Classification and clustering are important data mining tasks. What are the main differences between the 2 tasks?*

Clustering vs Classification

1. In clustering, you find similar groups of object/discover classes. In classification, you learn a model that predicts a class. That is, you learn to classify examples with respect to a prior given class structure.
2. Clustering is an unsupervised algorithm, and classification is a supervised algorithm (usually).
3. In clustering, data sets consists of attributes. In classification, data sets consists of attributes **and** class labels.
4. Similarity assessment which derives a distance function is critical in clustering. On the other hand, classifiers are learnt from set of classified examples (with classifiers).

Other answer might deserve credit! No more than 4 points total!

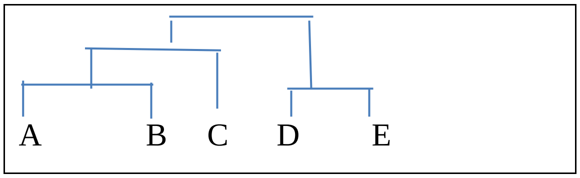
10. Hierarchical Clustering

Directions: *A dataset consisting of object A, B, C, D, E with the following distance matrix is given:*

Distance	A	B	C	D	E
A	0	1	2	8	9
B	0	0	3	6	5
C	0	0	0	7	10
D	0	0	0	0	4
E	0	0	0	0	0

a. *Assume single link hierarchical clustering is applied to the dataset? What dendrogram will be returned?*

The following Dendrogram would be returned:



b. *What is the goal of hierarchical clustering? How is it different from traditional clustering algorithms, such as K-Means and DBSCAN.*

The goal is to identify groups of similar objects incrementally (agglomerative/ divisive), which seeks to build a **hierarchy of clusters**, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. **Do not mention hierarchy.**

11. Data Visualization

Directions: *Answer the following questions/parts:*

(a) What are the different types of goals of data visualization?

- To explore data since we don't know any relationships.
- To analyze data and observe hypotheses.
- To verify or falsify existing information.
- To *present everything* known about the data.

(b) Explain what the following C.R.A.P. Design Principles emphasize:

They emphasize **contrast** and **repetition**.

Contrast is all about making distinct elements stand out. Contrast is used to drive a user's attention to specific elements in a design.

Repetition is how you maintain consistency in a design. It helps users familiarize with the way information is presented to them.

(c) 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?

The ratio between the width and the height of a rectangle is called the *aspect ratio*. The challenge when visualizing line charts is choosing the proper aspect ratio so that the chart appears correctly on all screens and surfaces it is presented on. We should consider the available space and required formats to fit the display (such as the width of a given column in an IEEE). Other considerations including banking at 45° and data ink ratio.

12. Non – parametric density estimation

a. Directions: *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.*

The general formula is given as:

$$\phi(v) = \frac{1}{n \cdot 2\pi \cdot h^2} \sum_i \exp\left(-\frac{(o.x - v.x)^2 - (o.y - v.y)^2}{2h^2}\right)$$

For our case, we have:

$$\phi_d(3,3) = \frac{1}{3 \cdot 2\pi}(e^{-9} + e^{-1} + e^{-5})$$

b. Directions: *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?*

The bandwidth determines how quickly the influence of a point (in the dataset) on the query point decreases with the increase of distance. h high few hills smooth display, h low rugged display with a lot variation/many hills. That is, h high captures regional variation and h low captures local variation.