Christoph F. Eick

Review1 COSC 3337 Fall 2019

September 30, 2019

1. Explain what the following C.R.A.P. Design Principles emphasize:
2. Alignment

With alignment, no element in a design is positioned arbitrarily. Each element visually connects to other elements, leading to cohesiveness in a design.

1. Proximity

The principle of proximity states that elements that are associated with each other should be placed closely (grouped together), and vice versa.

2. Supervised Learning & Decision Trees

1. What role do test sets play in supervised learning?

….to determine the accuracy of the learnt model

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

(12,4,6) (3,3,0)

( 9,1,0)

(0,0,6)

G(6/11,2/11,3/11) – (6/22\*G(0.5,0.5,0) + 10/22\* G(0.9,0,1,0) + 0)=

(1- (6/11)\*\*2-(3/11)\*\*2-(2/11\*\*2)- (6/22\*0.5)- 10/22\*(1-0.9\*\*2-0.1\*\*2)=

(121-36-9-4)/121 - …=

72/121=0.595-=

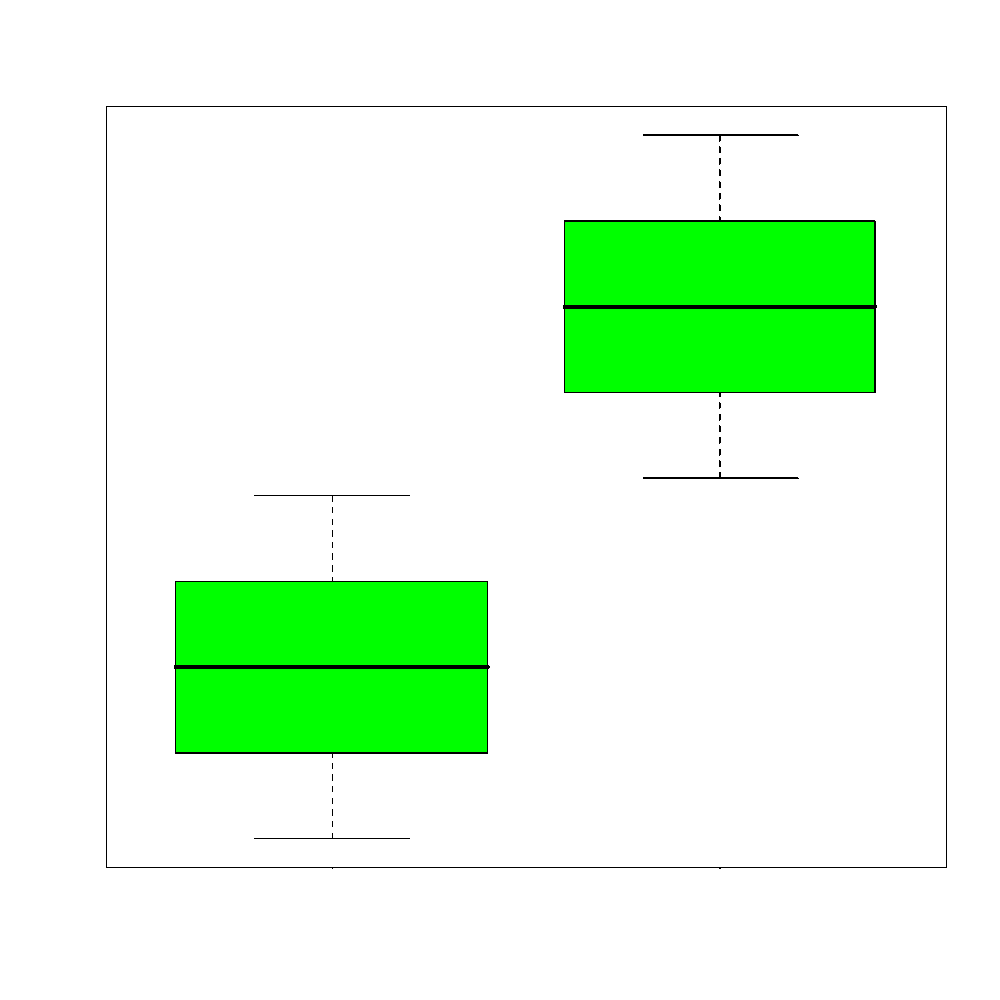
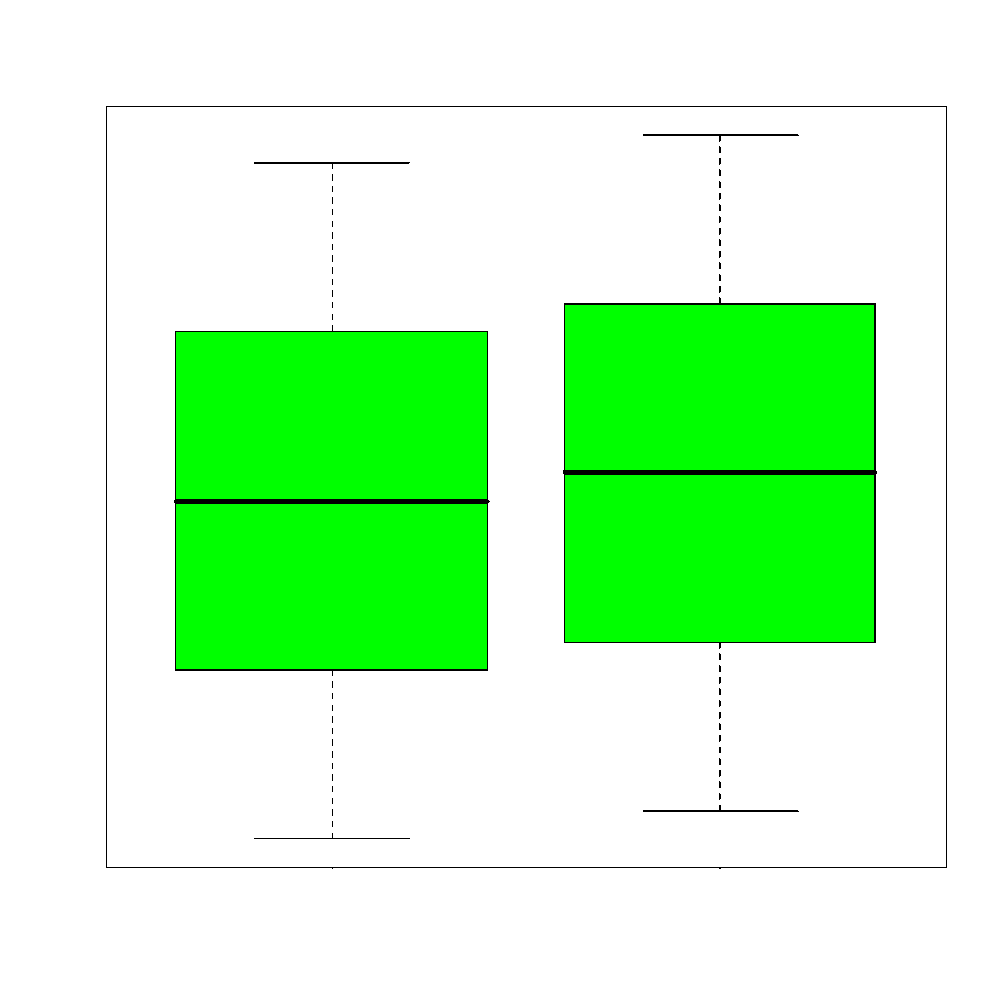
1. What does the size of the box of a boxplot measure; what statistical measure is it related to?

*The difference between the 25thand 75th quantile, also called IQR, of the attribute; the size of the box is used as an estimator of the standard deviation of the attribute.*

1. A R-boxplot (also called Turkey boxplots) of an attribute A has whiskers at 2 and 10; what does this tell you about attribute A? What attribute values are typically considered to be outliers in boxplots?

*The largest attribute value that is not an outlier is 10, and the smallest attribute value that is not an outlier is 2; all attribute values that are 1.5IQR or more above the 25% quantile or 1.5IQR below the the 75% quantile of the attribute are considered outliers.*

1. Comparing Histograms

Dramatically different: one is almost always higher (with exception of outliers); not much difference; one is slightly highter, but their boxes mostly overlap!

1. What is the main difference between ordinal and a nominal attributes?

*The values of ordinal attributes are ordered; this fact has to be considered when assessing similarity between two attribute values!*

1. What role does exploratory data analysis play in a data mining project?

**create background knowledge about the dataset and the task at hand [1], assess difficulty [1], provide knowledge to help select appropriate tools for the task[1], assess quality of data [1], validate data [1], help to form hypothesis [1], find issues, patterns and errors in data [1]**

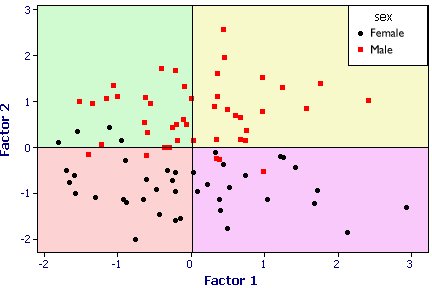
1. Assuming use 2D non parametric density estimation for a dateset D={(0,0), (2,2), (0,1)} and h=1. Give the formula that computes the density in the query points (1,1)---give the influences of each point!

ϕO(v) (General Formula)

ϕD(1,1))= (1/ (3\*2π))(e-1+e-1 + e-0.5)=… (YOU WERE ONLY ASKED TO GIVE THE Formula)

ϕD(2,2))= (1/ (3\*2π))(e-4+1 + e-2.5)=… (YOU WERE ONLY ASKED TO GIVE THE Formula)

1. Interpret the supervised scatter plot depicted below; moreover, assess the difficulty of separating males from females using Factor 1 / Factor 2 based on the scatter plot! [5]

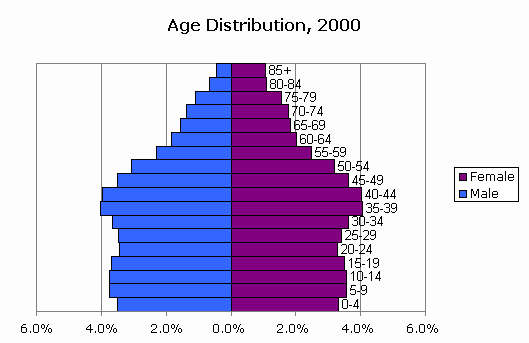


Both the female and male class have a uni-modal distribution; no gaps in data density are visible. Factor2 does mostly a good job in separating females and males; there is only overlap close to 0; Factor1 does a poor job separating the 2 classes. The classification task should not be too difficult as the examples are well separated although there are a few exceptions.

1. What is (are) the characteristic(s) of a good histogram (for an attribute)?

*It captures the most important characteristics of the underlying density function*

1. Interpret the following 2 histograms and their relationships which describe the male and female age distribution in the US, based on Census Data. Already Discussed in the Lecture!



**Both histograms: curves are continuous with no gabs or outliers, 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.

1. Assume you find out that two attributes have a correlation of 0.02; what does this tell you about the relationship of the two attributes? Answer the same question assuming the correlation is -0.98!

*0.02:= no linear relationships exists between the two attributes—but other relationships might exist; 0.98:=a strong linear relationship exists—if the value of one attribute goes up the value of the other goes down*

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