Christoph F. Eick

Review1 COSC 3337 Fall 2019[[1]](#footnote-1)

September 30, 2019

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

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[[2]](#footnote-2) for the following decision tree split (just giving the formula is fine!)

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

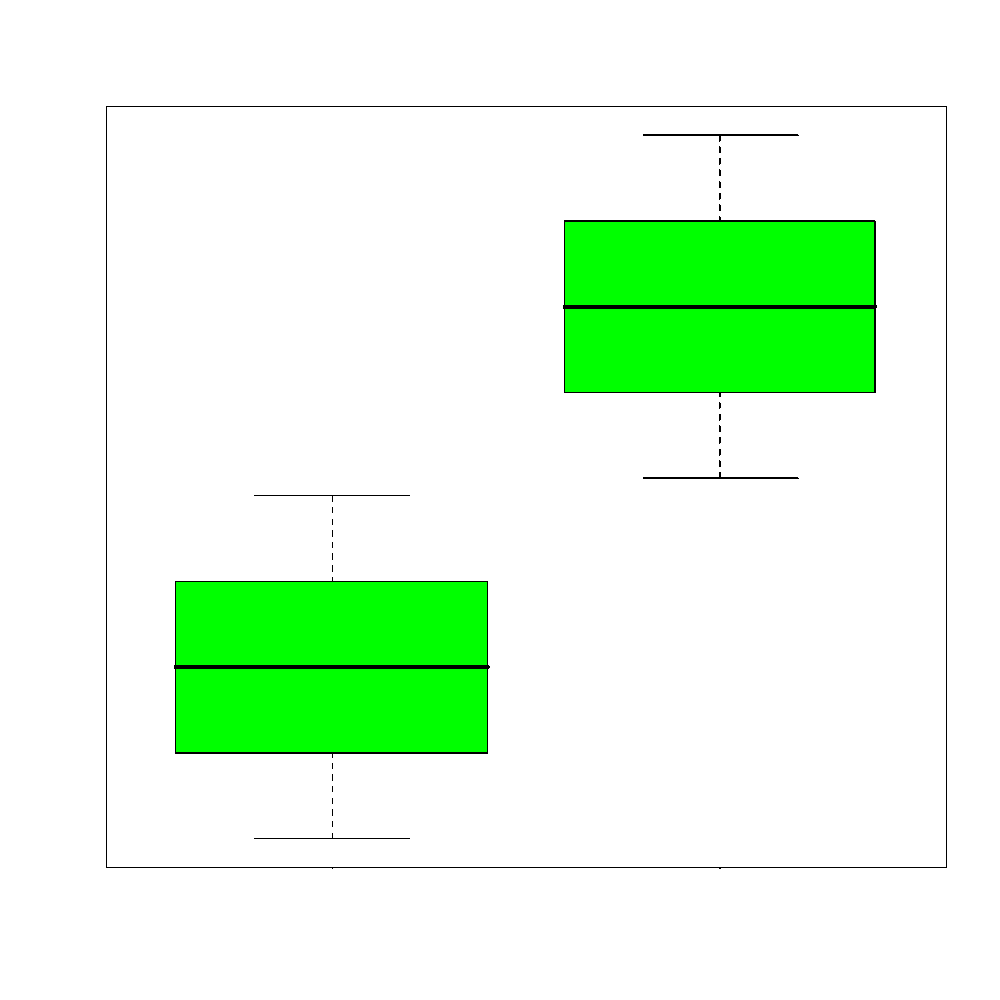
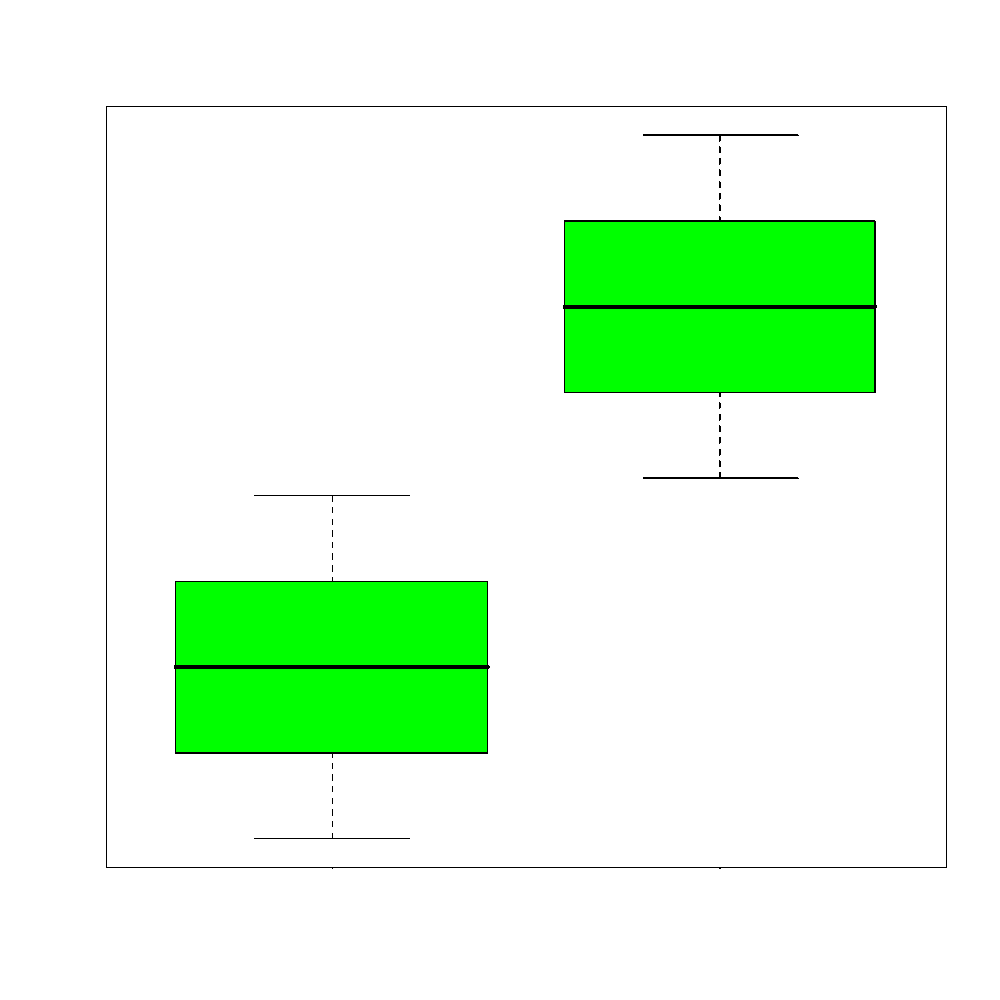
( 9,1,0)

(0,0,6)

1. What test does the Decision Tree Induction Algorithm prefer to create?

Splits that in the examples in subsequent nodes mostly belong/all belong to the same class

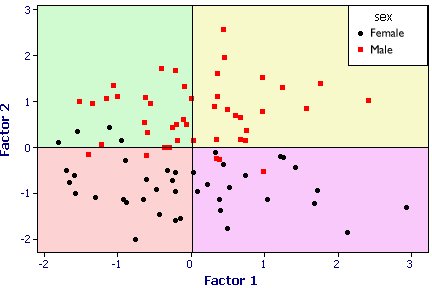
1. What does the size of the box of a boxplot measure; what statistical measure is it related to?
2. A R-boxplot (also called Turkey boxplots) of an attribute A has whiskers at 2 and 10; what does this tell you about attribute A? What attribute values are typically considered to be outliers in boxplots?
3. Comparing Histograms

1. What is the main difference between ordinal and a nominal attributes?
2. What role does exploratory data analysis play in a data mining project?
3. 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)!

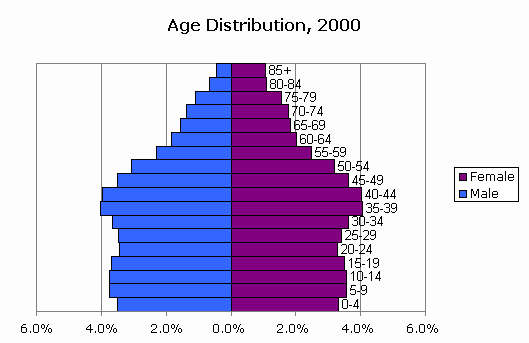
ϕO(v)

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)?
2. 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!

1. A subset of the problems 1-8 and 15 will be discussed on September 24, and a subset of problems 9-14 will be discussed on the lecture on September 26. [↑](#footnote-ref-1)
2. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-2)