Homework 1

Celine Mol and Emeric Szaboky

April 16, 2017

Note: *If your compiled pdf does not look right, number of spaces in* `indent1` *variable*

---

1. Discuss whether or not each of the following activities is a data mining task. (Answer YES/NO and explain your reasons.) [12]
   (a) Looking up customers of a company according to their profitability. **NO. There is no processing of data to gain insight unless you do something specific with the customers of a company.**
   (b) Computing the total sales of a company. **NO. This is just doing math with explicit data, there is no interpretation of insights.**
   (c) Predicting the future stock price of a company using historical records. **YES. This is a predictive regression data mining task.**
   (d) Sorting a student database based on student identification numbers. **NO. This is a database query.**
   (e) Predicting the outcomes of tossing a (fair) pair of dice. **YES. This could be considered a predictive classification data mining task because you can classify each toss in an existing category (2-12). This is a probability of a calculation.**
   (f) Extracting the frequencies of a sound wave. **YES. This is a descriptive data mining task for record data. This is signal processing.**

Predicting Algae Blooms[1] [38]

Background High concentrations of certain harmful algae in rivers constitute a serious ecological problem with a strong impact not only on river lifeforms, but also on water quality. Being able to monitor and perform an early forecast of algae blooms is essential to improving the quality of rivers.

With the goal of addressing this prediction problem, several water samples were collected in different European rivers at different times during a period of approximately 1 year. For each water sample, different chemical properties were measured as well as the frequency of occurrence of seven harmful algae. Some other characteristics of the water collection process were also stored, such as the season of the year, the river size, and the river speed.

Goal We want to understand how these frequencies are related to certain chemical attributes of water samples as well as other characteristics of the samples (like season of the year, type of river, etc.)

Data Description The data set consists of data for 200 water samples and each observation in the available datasets is in effect an aggregation of several water samples collected from the same river over a period of 3 months, during the same season of the year.Each observation contains information on 11 variables. Three of these variables are nominal and describe the season of the year when the water samples to be aggregated were collected, as well as the size and speed of the river in question. The eight remaining variables are values of different chemical parameters measured in the water samples forming the aggregation, namely: Maximum pH value, Minimum value of O_2 (oxygen), Mean value of Cl (chloride), Mean value of NO_3^- (nitrates), Mean value of NH_4^+ (ammonium), Mean of PO_4^3 (orthophosphate), Mean of total PO_4 (phosphate) and Mean of chlorophyll.

Associated with each of these parameters are seven frequency numbers of different harmful algae found in the respective water samples. No information is given regarding the names of the algae that were identified.

We can start the analysis by loading into R the data from the "algaeBloom.txt" file (the training data, i.e. the data that will be used to obtain the predictive models). To read the data from the file it is sufficient to issue the following command:

[1] This case study will introduce you to some basic steps of data mining: data pre-processing, exploratory data analysis, and predictive model construction throughout the quarter.

```
algae <- read.table('algaeBloom.txt',header=F,dec='.',
                    col.names=c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4',
                                'PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
                    na.strings=c('XXXXXXX'))
attach(algae)
head(algae,3)
```

```
##   season  size  speed mxPH mnO2    Cl   NO3     NH4    oPO4     PO4 Chla
## 1 winter small medium 8.00  9.8 60.80 6.238 578.000 105.000 170.000 50.0
## 2 spring small medium 8.35  8.0 57.75 1.288 370.000 428.750 558.750  1.3
## 3 autumn small medium 8.10 11.4 40.02 5.330 346.667 125.667 187.057 15.6
##    a1   a2  a3  a4   a5  a6  a7
## 1 0.0  0.0 0.0 0.0 34.2 8.3 0.0
## 2 1.4  7.6 4.8 1.9  6.7 0.0 2.1
## 3 3.3 53.6 1.9 0.0  0.0 0.0 9.7
```

2. ***Descriptive summary statistics*** Given the lack of further information on the problem domain, it is wise to investigate some of the statistical properties of the data, so as to get a better grasp of the problem. It is always a good idea to start our analysis with some kind of exploratory data analysis. A first idea of the statistical properties of the data can be obtained through a summary of its descriptive statistics. [17]

   (a) Count the number of observations for each season by using `summarise()` in `dplyr`. [5]

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
algae %>%
  group_by(season) %>%
  summarise(length(season))
```

```
## # A tibble: 4 x 2
##   season length(season)
##   <fctr>          <int>
## 1 autumn             40
## 2 spring             53
## 3 summer             45
## 4 winter             62
```

   (b) Are there missing values? Calculate the mean and variance of each chemical (Ignore a_1 through a_7). What do you notice about the magnitude of the two quantities for different chemicals? [6]

```
apply(algae[4:11], 2, mean, na.rm=T)   # na.rm=T indicates to R to remove missing values
```

```
##       mxPH       mnO2         Cl        NO3        NH4       oPO4
##   8.011734   9.117778  43.636279   3.282389 501.295828  73.590596
##        PO4       Chla
## 137.882101  13.971197
```

```
apply(algae[4:11], 2, var, na.rm=T)
```

```
##         mxPH         mnO2           Cl          NO3          NH4
```

```
## 3.579693e-01 5.718089e+00 2.193172e+03 1.426176e+01 3.851585e+06
##         oPO4          PO4         Chla
## 8.305850e+03 1.663938e+04 4.200827e+02
```

Yes, there are missing values, because if you calculate the mean and variance without `na.rm=T`, you get an NA value for all variables. This can also be verified by scrolling through the complete algae dataset and searching for the missing values, labeled NA. I notice that the variance of mxPH is close to 0, and the variances for NH4, PO4, oPO4, Cl, and Chla are the greatest, in respective order (greatest to least). I notice that the means for mxPH, mnO2, and NO3 are all in the ones place in size, while the means for Cl, oPO4, and Chla are in the tens place and the means for NH4 and PO4 are in the hundreds place.

(c) Mean and Variance is one measure of central tendency and spread of data. Median and Median Absolute Deviation are alternative measures of central tendency and spread.

For a univariate data set $X_1, X_2, ..., X_n$, the Median Absolute Deviation (MAD) is defined as the median of the absolute deviations from the data's median:

$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$$

Compute median and MAD of each chemical and compare the two sets of quantities (i.e., mean & variance vs. median & MAD). What do you notice? [6]

```
# na.rm=T indicates to R to remove missing values
apply(algae[4:11], 2, median, na.rm=T)  # median
```

```
##     mxPH     mnO2       Cl      NO3      NH4     oPO4      PO4     Chla
##   8.0600   9.8000  32.7300   2.6750 103.1665  40.1500 103.2855   5.4750
```

```
apply(algae[4:11], 2, mad, na.rm=T)     # Median Absolute Deviation (MAD)
```

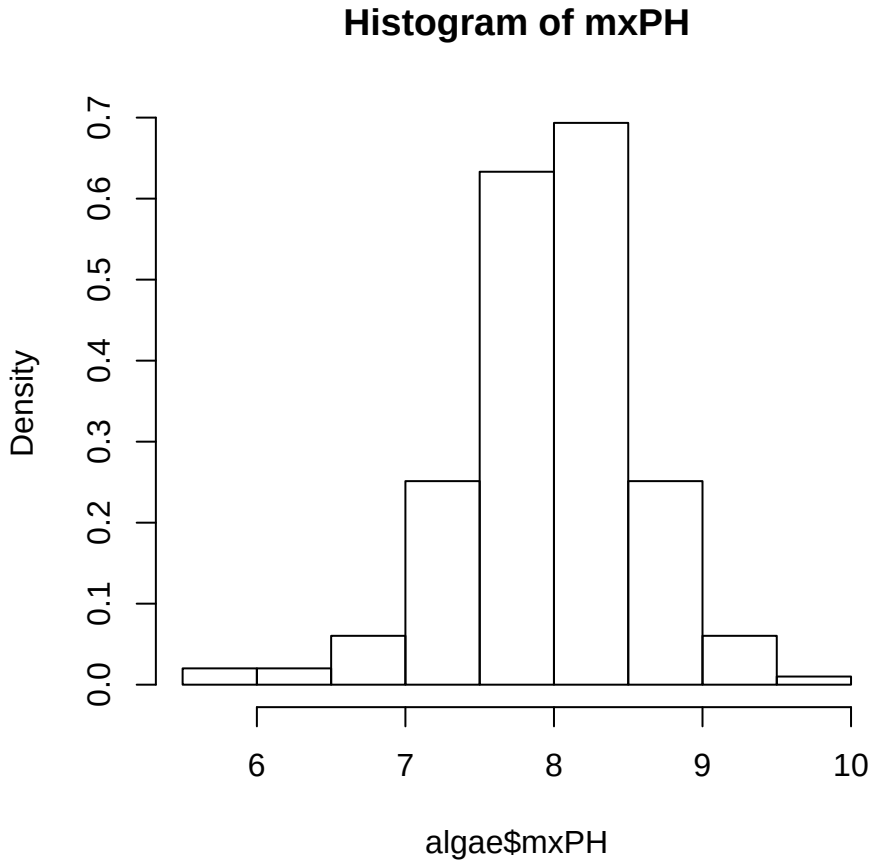
```
##       mxPH       mnO2         Cl        NO3        NH4       oPO4
##   0.504084   2.053401  33.249529   2.172009 111.617548  44.045822
##        PO4       Chla
## 122.321172   6.671700
```

I notice that for some values, such as mxPH, NO3 and mnO2, the median is very close to the mean, but for other values, such as Cl, NH4, oPO4, PO4, and Chla, the value of the median strays quite far from the mean, suggesting that there could be a few outliers present in the data, which is also represented by the high variance in the previous calculations. When you look at the Median Absolute Deviation (MAD), for values like mxPH, NO3, and mnO2, the absolute deviation from the median is quite low, but for values such as Cl, NH4, oPO4, PO4, and Chla, the absolute deviation from the median is quite high, which speaks to support our previous findings of the median straying far from the mean and the variance being high. We can see that the problematic variables Cl, NH4, oPO4, PO4, and Chla are constant for both measures of central tendency and spread of data (mean and variance & median and MAD). These variables have high variance, medians which stray vastly from their means, and large MAD values. These summary statistics suggest either the existence of outliers or less evenly/normally distributed data for these variables (i.e. abnormally heavy tails of distributions).

3. ***Data visualization*** Most of the time, the information in the data set is also well captured graphically. Histogram, scatter plot, boxplot, Q-Q plot are frequently used tools for data visualization. [21]

(a) Produce a histogram of *mxPH* with the title 'Histogram of mxPH' based on algae data set. Use an appropriate argument to show the probability instead of the frequency as the vertical axis. (Hint: read the help file for function `hist()`). Is the distribution skewed? [4]

```
hist(algae$mxPH, freq=FALSE, main='Histogram of mxPH')
```

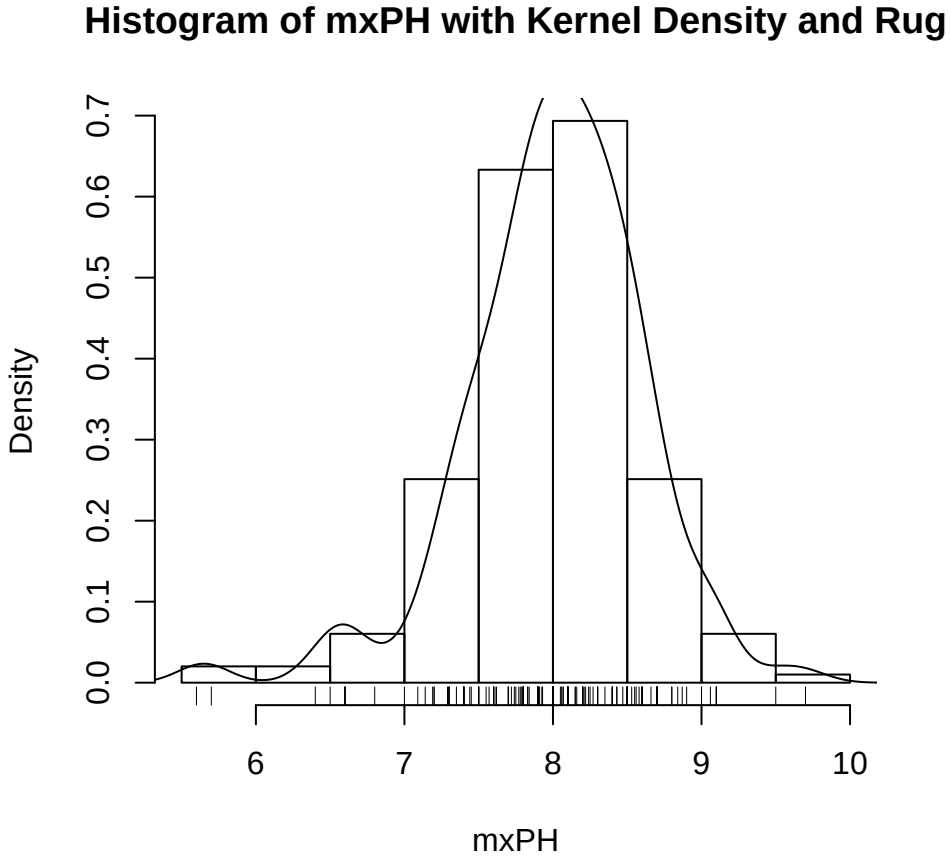



```
# freq=FALSE to specify probability densities on the vertical axis
```

The distribution of mxPH does not look significantly skewed. It has a longer left tail and is very slightly skewed to the left (Negative Skewness).

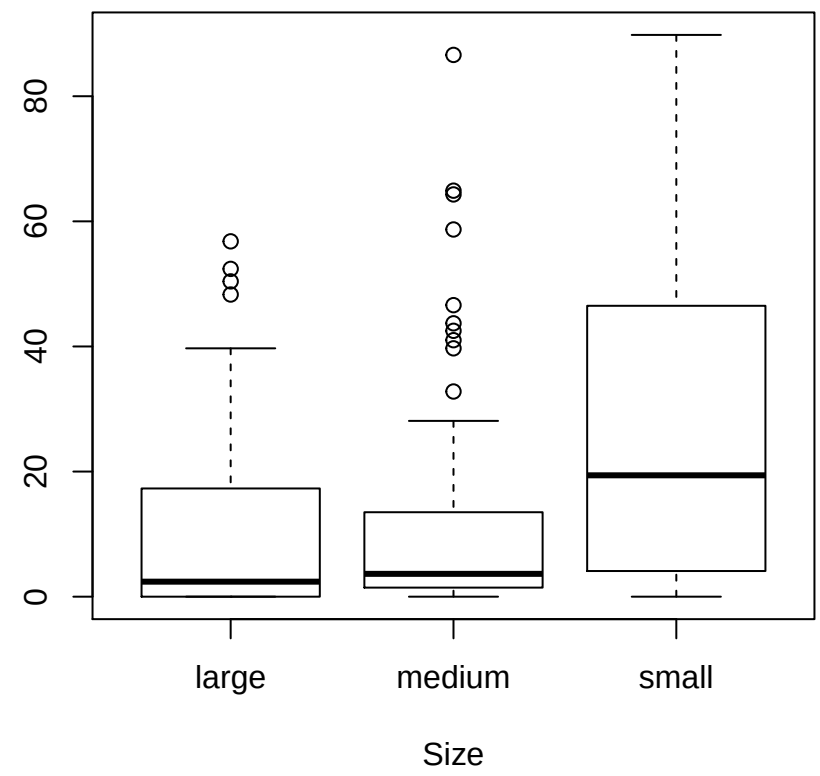
(b) Add a density curve using `density()` and rug plots using `rug()`to above histogram. [4]

```
with(algae, {
  hist(mxPH, freq=FALSE, main='Histogram of mxPH with Kernel Density and Rug')
  lines(density(mxPH, na.rm=T))
  rug(mxPH)
})
```



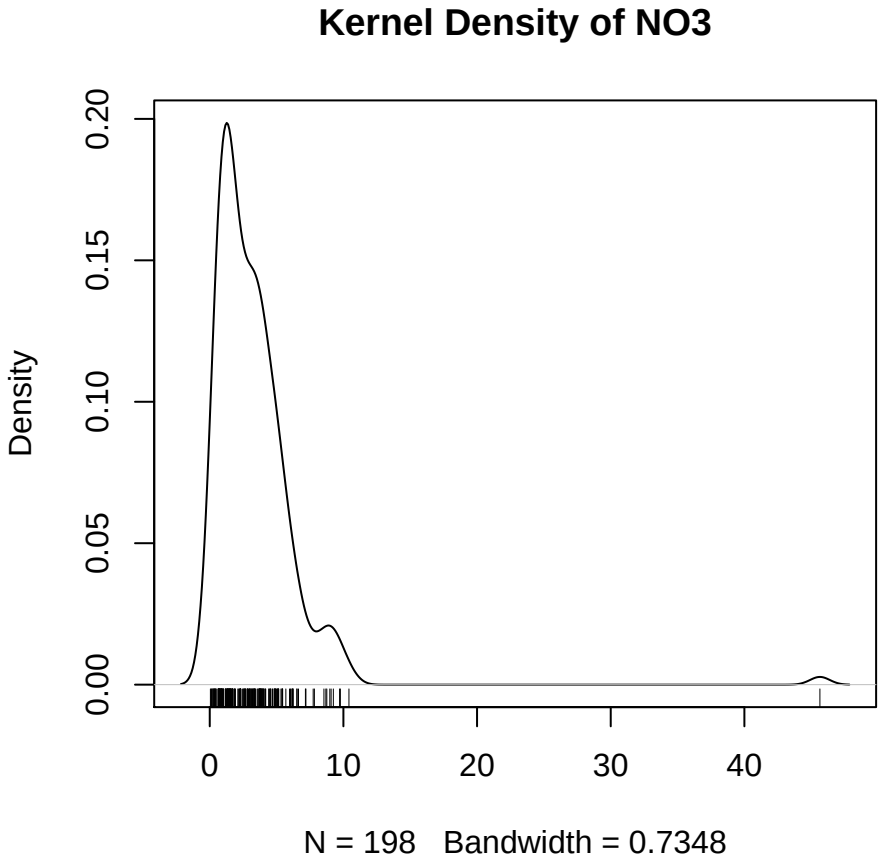
(c) Create a boxplot with the title 'A Conditioned Boxplot of Algal a_1' for a_1 grouped by *size*. (Refer to help page for `boxplot()` on using formula notation). [4]

```
boxplot(algae$a1~algae$size, xlab="Size", main='A Conditioned Boxplot of Algal a1 grouped by Size'
```



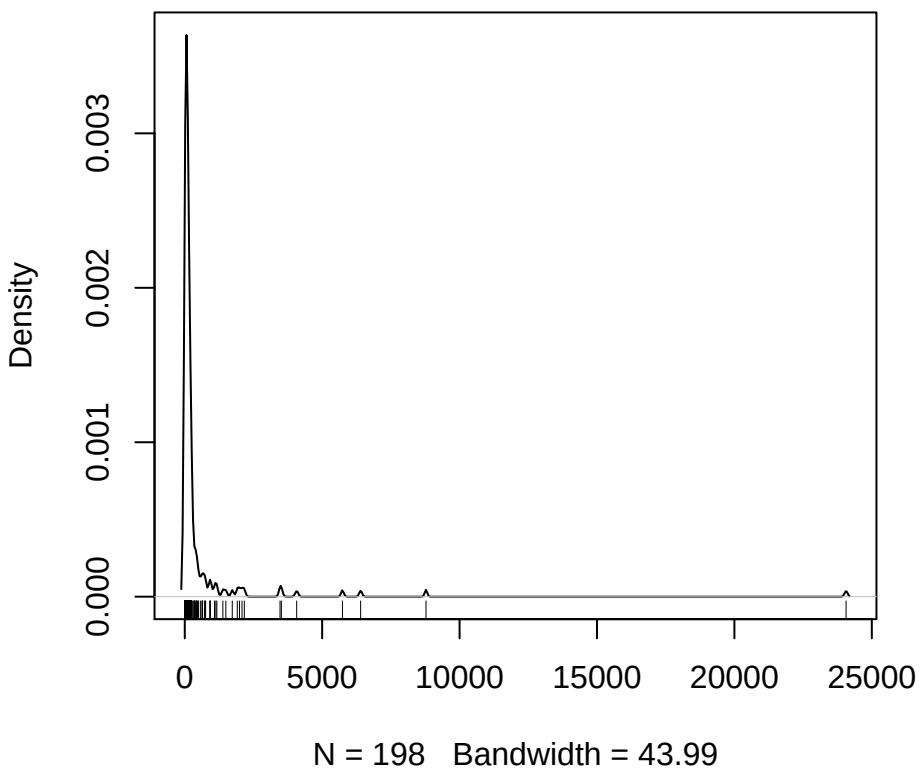
(d) Are there any outliers for $NO3$ and $NH4$? How many observations would you consider as outliers? How did you arrive at this conclusion? [5]

```
with(algae, {
  plot(density(NO3, na.rm=T), main="Kernel Density of NO3")
  rug(NO3)
})
```



Based on looking at this density function of N03, we can see that there is one outlier on the right that is not consistent with the mean or spread of the data.

```
with(algae, {
  plot(density(NH4, na.rm=T), main="Kernel Density of NO4")
  rug(NH4)
})
```



As we can see as well with the density function of NH4, this chemical also contains a few outliers at various spreads, four significantly visible ones on the rug and a few more. You can see this because the data is skewed to the right, in addition to seeing the lines on the rug.

(e) Compare mean & variance vs. median & MAD for $NO3$ and $NH4$. What do you notice? Can you conclude which set of measures is more robust when outliers are present? [4]

```
apply(algae[7:8], 2, mean, na.rm=T)    # mean
```

```
##        NO3        NH4
##   3.282389 501.295828
```

```
apply(algae[7:8], 2, var, na.rm=T)     # variance
```

```
##          NO3          NH4
## 1.426176e+01 3.851585e+06
```

```
apply(algae[7:8], 2, median, na.rm=T)  # median
```

```
##      NO3      NH4
##   2.6750 103.1665
```

```
apply(algae[7:8], 2, mad, na.rm=T)     # Median Absolute Deviation (MAD)
```

```
##        NO3        NH4
##   2.172009 111.617548
```

```
# Means:     NO3~3.2824     NH4~501.2958
# Variances: NO3~14.262     NH4~3851585
# Medians:   NO3~2.675      NH4~103.1665
# MADs:      NO3~2.172      NH4~111.6175
```

As we can see above, for NO3, the median is quite close to the mean, and the Median Absolute Deviation is significantly smaller than the variance. For NH4, the median and the mean are quite different, with the

mean being much higher, and the Median Absolute Deviation is significantly smaller than the variance. Since our outliers were much greater for NO3 and NH4 than our other values, we can see that calculating the median and the MAD is more robust when outliers are present, since the mean and variance are clearly highly influenced by the outliers that are present.

Additional problems for 231 students

4. Prove L_∞ norm (Page 70, Tan) is supremum norm: i.e., $\|x\|_\infty = \lim_{r\to\infty} \left(\sum_{i=1}^{n} |x_i|^r \right)^{1/r} = \max_{1\le i\le n} |x_i|$. [10]

   `__See attachment.__`

5. Show that the following measures are distances by showing properties on Pages 70-71, Tan, are satisfied. [15]
   (a) $d(x, y) = \|x - y\|_2$ [7]
   (b) $d(x, y) = \|x - y\|_\infty$ [8]

   See attachment.