Homework 1

PSTAT 131/231, Fall 2021

Due on October 11, 2021 at 23:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

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:

```
algae <- read_table2("algaeBloom.txt", col_names=
                       c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4',
                         'oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
                     na="XXXXXXX")

glimpse(algae)
```

1. ***Descriptive summary statistics*** (10 pts in total) 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.

(a) (2 pts) Count the number of observations in each size using `summarise()` in `dplyr`.

(b) (1 pts) Are there missing values? (2 pts) Calculate the mean and variance of each chemical (Ignore a_1 through a_7). (1 pts) What do you notice about the magnitude of the two quantities for different chemicals?

(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:

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

(3 pts) Compute median and MAD of each chemical and compare the two sets of quantities (i.e., mean & variance vs. median & MAD). (1 pts) What do you notice?

2. ***Data visualization*** (8 pts in total) 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. Use ggplot for all of these visualizations.

(a) (2 pts) Produce a histogram of $mnO2$ with the title 'Histogram of mnO2' based on algae data set. (1 pts) Use an appropriate argument to show the probability instead of the frequency as the vertical axis. (Hint: look at the examples in the help file for function `geom_histogram()`). (1 pts) Is the distribution skewed?

(b) (1 pts) Add a density curve using `geom_density()` and (1 pts) rug plots using `geom_rug()` to above histogram.

(c) (1 pts) Create a boxplot with the title 'A conditioned Boxplot of Algal a_3' for a_3 grouped by *speed*. (Refer to help page for `geom_boxplot()`). (1 pts) What do you notice?

3. ***Dealing with missing values*** (8 pts in total)

(a) (2 pts) How many observations contain missing values? (2 pts) How many missing values are there in each variable?

(b) (3 pts) **Removing observations with missing values**: use `filter()` function in `dplyr` package to observations with any missing value, and save the resulting dataset (without missing values) as `algae.del`. (1 pts) Report how many observations are in `algae.del`.

Hint: `complete.cases()` may be useful.

4. In lecture we present the bias-variance tradeoff that takes the form

$$\text{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \text{Var}(\hat{f}(\mathbf{x}_0)) + \left[\text{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \text{Var}(\varepsilon),$$

where the underlying model $Y = f(X) + \varepsilon$ satisifes: (1) ε is a zero-mean random noise, and X is non-random (all randomness in Y comes from ε); (2) $(\mathbf{x}_0, y_0)$ is a test observation, independent of the training set, and drawn from the same model; (3) $\hat{f}(\cdot)$ is the estimate of f obtained on a training set.

(a) (2 pts) Which of the term(s) in the bias-variance tradeoff above represent the reducible error? (2 pts) Which term(s) represent the irreducible error?

(b) (4 pts) Use the bias-variance tradeoff above to show that the expected test error is always at least as large as the irreducible error.

---

5. **(231 Only)** (6 pts) Prove the bias-variance tradeoff, where $\text{Bias}(\hat{f}(\mathbf{x}_0)) = \text{E}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0)$. Hint: reogranize terms in the expected test error by adding and subtracting $\text{E}[\hat{f}(\mathbf{x}_0)]$.

6. **(231 Only)** Distance metrics are a very important concept used in kNN. A distance metric has to satisfy the following properties:

- *Positivity*:
  - $d(x, y) \geq 0$
  - $d(x, y) = 0$ only if $x = y$
- *Symmetry*:
  - $d(x, y) = d(y, x)$ for all x and y
- *Triangle Inequality*:
  - $d(x, z) \leq d(x, y) + d(y, z)$ for $x,\ y,$ and z

Here, x, y are p-dimensional vectors. Show that the following measures are distance metrics by showing the above properties hold:

1. (3 pts) $d(x, y) = \|x - y\|_2$
2. (3 pts) $d(x, y) = \|x - y\|_\infty$