Assignment 0

CS 750/850 Machine Learning

Landon Buell

23 January 2020

- **Due**: Monday 1/27 at 11:59PM
- **Submisssion**: Turn in as a **PDF** and the **source code** (R,Rmd,py,ipynb) on MyCourses
- **Questions**: Piazza and Office hours: *Marek*: Wed 1:30-3:00pm, Soheil: Mon 2-4pm, Xihong: Thu 1:30-3:30pm
- **Extra credit**: Especially good questions or helpful answers on Piazza regarding the assignment earn up to 5 points extra credit towards the assignment grade.

Problem 1 [33%]

What are the advantages and disadvantages of very flexible (vs less flexible) approach for regression or classification?

1. When would be a more flexible approach preferable?
   Regression allows for a far more flexible approach to building a function $\hat{f}$. By using a set of input data, X, we can produce a (mostly) continuous function $Y = f(X)$ that can effectly cover a larger range of values than classification. This more flexible approach would be useful for predicting a set of continuous output. For example, If you wanted to produce a function that predicted a country's GDP based on it's own economic growth, and the growth of it's neighboring countries GDP, then a more flexible regression would be of great benefit.

2. What about a less-flexible approach? Classification allows for a less flexible approach to buidling the function $\hat{f}$. With the set of input data, X, then we can create a set of discrete bins to place our output values, Y into. For example. if we were trying to build a model that tries to determine a person age based on a picture of them, it would be far more valid to attempt to group them by (for example) sets of 5 years. Trying to produce a regression to determine a continuous age down to monthsm weeks and so forth would be far less helpful than a less flexible classifier.

Problem 2 [33%]

Install and learn to use R (https://www.r-project.org/) or Python, read the labs in Chapter 2 of the textbook. We recommend that you use R Notebooks of RStudio to typeset homeworks. Jupyter is a comparable tool for Python. Use Python or another tool (like MATLAB or Julia) if you have some experience and you will not need help from the TA/instructor. Then:

1. Download the advertising dataset (`Advertising.csv`) from http://www-bcf.usc.edu/~gareth/ISL/data.html and load it into R/Python (use function `read.csv()` in R or Pandas in Python)

```
filename = 'Advertising.csv'
filedata = read.csv(file=filename)
print(class(filedata))
```

```
## [1] "data.frame"
```

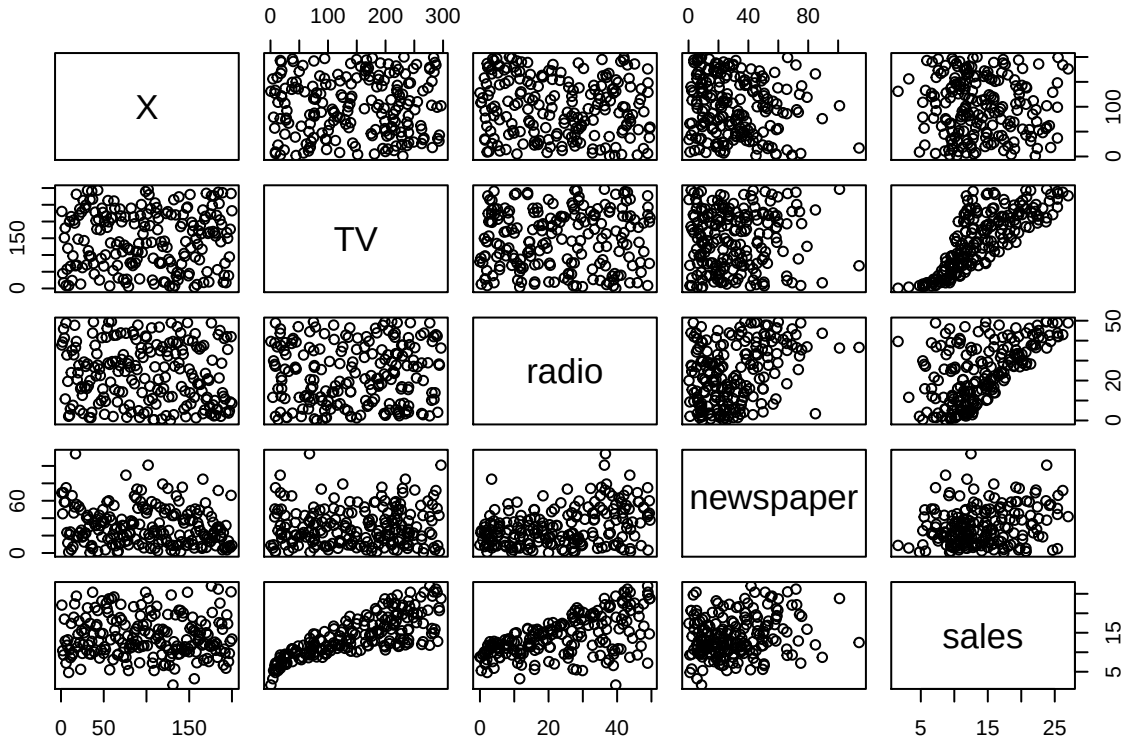
2. What are the minimum, maximum, and mean value of each feature? (in R use function `summary()` and or `range()`)

```
summary(filedata)
```

```
##        X                TV              radio          newspaper
##  Min.   :  1.00   Min.   :  0.70   Min.   : 0.000   Min.   :  0.30
##  1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75
##  Median :100.50   Median :149.75   Median :22.900   Median : 25.75
##  Mean   :100.50   Mean   :147.04   Mean   :23.264   Mean   : 30.55
##  3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10
##  Max.   :200.00   Max.   :296.40   Max.   :49.600   Max.   :114.00
##      sales
##  Min.   : 1.60
##  1st Qu.:10.38
##  Median :12.90
##  Mean   :14.02
##  3rd Qu.:17.40
##  Max.   :27.00
```

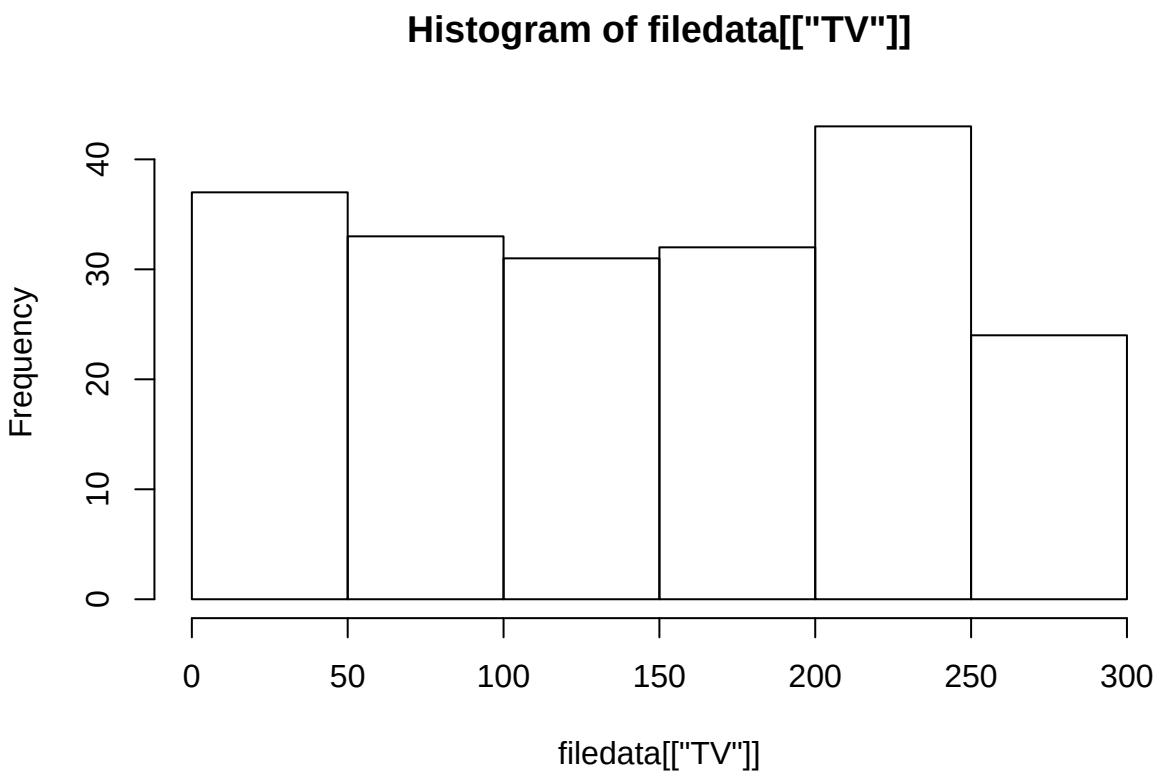
3. Produce a scatterplot matrix of all variables (in R use function `pairs()`)

```
pairs(filedata)
```



4. Produce a histogram of TV advertising (in R use function `hist()`)

```
hist(filedata[["TV"]])
```


Histogram of filedata[["TV"]]
Frequency
0
10
20
30
40
0
50
100
150
200
250
300
filedata[["TV"]]

Problem 3 [34%]

Describe some real-life applications for machine learning.

1. Describe one real-life application in which *classification* combined with *prediction* may be useful. Describe the response and predictors.
   In a self driving, car. A front- facing camera could use some sort of classification methods (SGD) to be able to indentify brake lights or trafic lights in an image. Once a brake light or no brake light or red traffic light is identified, another prediciton method could be used to predict when traffic will start up or start again.

2. Describe one real-life application in which *classification* combined with *inference* may be useful. Describe the response and predictors.
   Again, with self-driving cars (My brother has a Tesla) a camera classification algorithm could be use to identify an image with a car changing lanes in front of you. We can then use an infrence algorithm to detemine how the distance you and the car affects how strongly and soon the brakes need by applied.

3. Describe one real-life application in which *regression* combined with *prediction* may be useful. Describe the response and predictors.
   If we track the median salarays for Americans in the early 2000's, we can use regression to generate a model (or a function f) that tracks median salaries as a function of the year. We can then used this trained model to *predict* the possible median salaries for the next few years in the 2020's.

4. Describe one real-life application in which *regression* combined with *inference* may be useful. Describe the response and predictors.
   Using a similar set of data, we can use a regression algorithm to produce a function f. We can then use the features in that data set to infer how something like their geography affects either their median salary or their median salary's rate of change.

Optional Problem O3 [39%]

This problem can be substituted for Problem 3 above, for 5 points extra credit. At most one of the problems 3 and O3 will be considered.

Read sections 1.2, 1.2.1, 1.2.2 in [Bishop, C. M. (2006). Pattern Recognition and Machine Learning] and solve *Exercise 1.5* in the said textbook.

Hints

1. An easy way to launch help for any function in R, such as `summary`, is to execute: `> ?summary`
2. See http://rmarkdown.rstudio.com/pdf_document_format.html for how to generate a PDF from an R notebook in R-studio. You will also need to install LaTeXwhich you can get from https://www.latex-project.org/get/
3. For more advanced (and prettier?) plotting capabilities, see the package `ggplot`: http://ggplot2.tidyverse.org/ and https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf
4. If you think you may struggle with R, consider signing up for MATH 759, a 1-credit online introduction to R.