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?
2. What about a less-flexible approach?

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)

```r
print("Reading 'Advertising.csv' file")
```

```
## [1] "Reading 'Advertising.csv' file"
```

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

```
##        X    TV radio newspaper sales
## 1     1 230.1  37.8      69.2  22.1
## 2     2  44.5  39.3      45.1  10.4
## 3     3  17.2  45.9      69.3   9.3
## 4     4 151.5  41.3      58.5  18.5
## 5     5 180.8  10.8      58.4  12.9
## 6     6   8.7  48.9      75.0   7.2
## 7     7  57.5  32.8      23.5  11.8
## 8     8 120.2  19.6      11.6  13.2
## 9     9   8.6   2.1       1.0   4.8
## 10   10 199.8   2.6      21.2  10.6
## 11   11  66.1   5.8      24.2   8.6
## 12   12 214.7  24.0       4.0  17.4
## 13   13  23.8  35.1      65.9   9.2
## 14   14  97.5   7.6       7.2   9.7
## 15   15 204.1  32.9      46.0  19.0
## 16   16 195.4  47.7      52.9  22.4
## 17   17  67.8  36.6     114.0  12.5
## 18   18 281.4  39.6      55.8  24.4
## 19   19  69.2  20.5      18.3  11.3
## 20   20 147.3  23.9      19.1  14.6
## 21   21 218.4  27.7      53.4  18.0
## 22   22 237.4   5.1      23.5  12.5
## 23   23  13.2  15.9      49.6   5.6
## 24   24 228.3  16.9      26.2  15.5
## 25   25  62.3  12.6      18.3   9.7
## 26   26 262.9   3.5      19.5  12.0
## 27   27 142.9  29.3      12.6  15.0
## 28   28 240.1  16.7      22.9  15.9
## 29   29 248.8  27.1      22.9  18.9
## 30   30  70.6  16.0      40.8  10.5
## 31   31 292.9  28.3      43.2  21.4
## 32   32 112.9  17.4      38.6  11.9
## 33   33  97.2   1.5      30.0   9.6
## 34   34 265.6  20.0       0.3  17.4
## 35   35  95.7   1.4       7.4   9.5
## 36   36 290.7   4.1       8.5  12.8
## 37   37 266.9  43.8       5.0  25.4
## 38   38  74.7  49.4      45.7  14.7
## 39   39  43.1  26.7      35.1  10.1
## 40   40 228.0  37.7      32.0  21.5
## 41   41 202.5  22.3      31.6  16.6
## 42   42 177.0  33.4      38.7  17.1
## 43   43 293.6  27.7       1.8  20.7
## 44   44 206.9   8.4      26.4  12.9
## 45   45  25.1  25.7      43.3   8.5
## 46   46 175.1  22.5      31.5  14.9
## 47   47  89.7   9.9      35.7  10.6
## 48   48 239.9  41.5      18.5  23.2
## 49   49 227.2  15.8      49.9  14.8
## 50   50  66.9  11.7      36.8   9.7
## 51   51 199.8   3.1      34.6  11.4
## 52   52 100.4   9.6       3.6  10.7
## 53   53 216.4  41.7      39.6  22.6
## 54   54 182.6  46.2      58.7  21.2
## 55   55 262.7  28.8      15.9  20.2
## 56   56 198.9  49.4      60.0  23.7
## 57   57   7.3  28.1      41.4   5.5
## 58   58 136.2  19.2      16.6  13.2
## 59   59 210.8  49.6      37.7  23.8
## 60   60 210.7  29.5       9.3  18.4
## 61   61  53.5   2.0      21.4   8.1
## 62   62 261.3  42.7      54.7  24.2
## 63   63 239.3  15.5      27.3  15.7
## 64   64 102.7  29.6       8.4  14.0
## 65   65 131.1  42.8      28.9  18.0
## 66   66  69.0   9.3       0.9   9.3
## 67   67  31.5  24.6       2.2   9.5
## 68   68 139.3  14.5      10.2  13.4
## 69   69 237.4  27.5      11.0  18.9
## 70   70 216.8  43.9      27.2  22.3
## 71   71 199.1  30.6      38.7  18.3
## 72   72 109.8  14.3      31.7  12.4
## 73   73  26.8  33.0      19.3   8.8
## 74   74 129.4   5.7      31.3  11.0
## 75   75 213.4  24.6      13.1  17.0
## 76   76  16.9  43.7      89.4   8.7
## 77   77  27.5   1.6      20.7   6.9
## 78   78 120.5  28.5      14.2  14.2
## 79   79   5.4  29.9       9.4   5.3
## 80   80 116.0   7.7      23.1  11.0
## 81   81  76.4  26.7      22.3  11.8
## 82   82 239.8   4.1      36.9  12.3
## 83   83  75.3  20.3      32.5  11.3
## 84   84  68.4  44.5      35.6  13.6
## 85   85 213.5  43.0      33.8  21.7
## 86   86 193.2  18.4      65.7  15.2
## 87   87  76.3  27.5      16.0  12.0
## 88   88 110.7  40.6      63.2  16.0
## 89   89  88.3  25.5      73.4  12.9
## 90   90 109.8  47.8      51.4  16.7
## 91   91 134.3   4.9       9.3  11.2
## 92   92  28.6   1.5      33.0   7.3
## 93   93 217.7  33.5      59.0  19.4
## 94   94 250.9  36.5      72.3  22.2
## 95   95 107.4  14.0      10.9  11.5
## 96   96 163.3  31.6      52.9  16.9
## 97   97 197.6   3.5       5.9  11.7
## 98   98 184.9  21.0      22.0  15.5
## 99   99 289.7  42.3      51.2  25.4
## 100 100 135.2  41.7       45.9  17.2
## 101 101 222.4   4.3       49.8  11.7
## 102 102 296.4  36.3      100.9  23.8
## 103 103 280.2  10.1       21.4  14.8
## 104 104 187.9  17.2       17.9  14.7
## 105 105 238.2  34.3        5.3  20.7
## 106 106 137.9  46.4       59.0  19.2
## 107 107  25.0  11.0       29.7   7.2
## 108 108  90.4   0.3       23.2   8.7
## 109 109  13.1   0.4       25.6   5.3
## 110 110 255.4  26.9        5.5  19.8
## 111 111 225.8   8.2       56.5  13.4
## 112 112 241.7  38.0       23.2  21.8
## 113 113 175.7  15.4        2.4  14.1
## 114 114 209.6  20.6       10.7  15.9
## 115 115  78.2  46.8       34.5  14.6
## 116 116  75.1  35.0       52.7  12.6
## 117 117 139.2  14.3       25.6  12.2
## 118 118  76.4   0.8       14.8   9.4
## 119 119 125.7  36.9       79.2  15.9
## 120 120  19.4  16.0       22.3   6.6
## 121 121 141.3  26.8       46.2  15.5
## 122 122  18.8  21.7       50.4   7.0
## 123 123 224.0   2.4       15.6  11.6
## 124 124 123.1  34.6       12.4  15.2
## 125 125 229.5  32.3       74.2  19.7
## 126 126  87.2  11.8       25.9  10.6
## 127 127   7.8  38.9       50.6   6.6
## 128 128  80.2   0.0        9.2   8.8
## 129 129 220.3  49.0        3.2  24.7
## 130 130  59.6  12.0       43.1   9.7
## 131 131   0.7  39.6        8.7   1.6
## 132 132 265.2   2.9       43.0  12.7
## 133 133   8.4  27.2        2.1   5.7
## 134 134 219.8  33.5       45.1  19.6
## 135 135  36.9  38.6       65.6  10.8
## 136 136  48.3  47.0        8.5  11.6
## 137 137  25.6  39.0        9.3   9.5
## 138 138 273.7  28.9       59.7  20.8
## 139 139  43.0  25.9       20.5   9.6
## 140 140 184.9  43.9        1.7  20.7
## 141 141  73.4  17.0       12.9  10.9
## 142 142 193.7  35.4       75.6  19.2
## 143 143 220.5  33.2       37.9  20.1
## 144 144 104.6   5.7       34.4  10.4
## 145 145  96.2  14.8       38.9  11.4
## 146 146 140.3   1.9        9.0  10.3
## 147 147 240.1   7.3        8.7  13.2
## 148 148 243.2  49.0       44.3  25.4
## 149 149  38.0  40.3       11.9  10.9
## 150 150  44.7  25.8       20.6  10.1
## 151 151 280.7  13.9       37.0  16.1
## 152 152 121.0   8.4       48.7  11.6
## 153 153 197.6  23.3       14.2  16.6
## 154 154 171.3  39.7       37.7  19.0
## 155 155 187.8  21.1        9.5  15.6
## 156 156   4.1  11.6        5.7   3.2
## 157 157  93.9  43.5       50.5  15.3
## 158 158 149.8   1.3       24.3  10.1
## 159 159  11.7  36.9       45.2   7.3
## 160 160 131.7  18.4       34.6  12.9
## 161 161 172.5  18.1       30.7  14.4
## 162 162  85.7  35.8       49.3  13.3
## 163 163 188.4  18.1       25.6  14.9
## 164 164 163.5  36.8        7.4  18.0
## 165 165 117.2  14.7        5.4  11.9
## 166 166 234.5   3.4       84.8  11.9
## 167 167  17.9  37.6       21.6   8.0
## 168 168 206.8   5.2       19.4  12.2
## 169 169 215.4  23.6       57.6  17.1
## 170 170 284.3  10.6        6.4  15.0
## 171 171  50.0  11.6       18.4   8.4
## 172 172 164.5  20.9       47.4  14.5
## 173 173  19.6  20.1       17.0   7.6
## 174 174 168.4   7.1       12.8  11.7
## 175 175 222.4   3.4       13.1  11.5
## 176 176 276.9  48.9       41.8  27.0
## 177 177 248.4  30.2       20.3  20.2
## 178 178 170.2   7.8       35.2  11.7
## 179 179 276.7   2.3       23.7  11.8
## 180 180 165.6  10.0       17.6  12.6
## 181 181 156.6   2.6        8.3  10.5
## 182 182 218.5   5.4       27.4  12.2
## 183 183  56.2   5.7       29.7   8.7
## 184 184 287.6  43.0       71.8  26.2
## 185 185 253.8  21.3       30.0  17.6
## 186 186 205.0  45.1       19.6  22.6
## 187 187 139.5   2.1       26.6  10.3
## 188 188 191.1  28.7       18.2  17.3
## 189 189 286.0  13.9        3.7  15.9
## 190 190  18.7  12.1       23.4   6.7
## 191 191  39.5  41.1        5.8  10.8
## 192 192  75.5  10.8        6.0   9.9
## 193 193  17.2   4.1       31.6   5.9
## 194 194 166.8  42.0        3.6  19.6
## 195 195 149.7  35.6        6.0  17.3
## 196 196  38.2   3.7       13.8   7.6
## 197 197  94.2   4.9        8.1   9.7
## 198 198 177.0   9.3        6.4  12.8
## 199 199 283.6  42.0       66.2  25.5
## 200 200 232.1   8.6        8.7  13.4
```

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

```
print("Summary of 'Advertising.csv'")
```

```
## [1] "Summary of 'Advertising.csv'"
```

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

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.
2. Describe one real-life application in which *classification* combined with *inference* may be useful. Describe the response and predictors.
3. Describe one real-life application in which *regression* combined with *prediction* may be useful. Describe the response and predictors.
4. Describe one real-life application in which *regression* combined with *inference* may be useful. Describe the response and predictors.

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.