Problem Set 1

Sixue Liu

January 17, 2020

Statistical and Machine Learning

Supervised learning and unsupervised learning are two distinct machine learning techniques. The major difference in supervised learning and unsupervised learning is if the data set has label Y.

In supervised learning, we have distinct relationship between X and Y. Therefore, we are more interested in detecting the difference between our predicted Y and the true Y. For this aim, the reason to split the data set into training set, validation set, and test set is to help us train better model with more similar prediction results. That is to say, the target we are interested in is getting accurate predictions in the testing data set. Also, we will select different models based on different Y. If Y is a continuous variable, it's more likely to use regression like linear regression, ridge regression, LASSO, etc. If Y is a label variable, we can use logistic regression or other classification methods like decision tree, random forest, K nearest neighbor, and Support Vector Machines, etc.

In unsupervised learning, there is no Y we could have as a target variable. What we have is only feature matrix of X. The target we are interested in is cluster the samples by their features. What we are doing with unlabeled data is very like predicting categories those samples should belong to. I think there is a big difference between unsupervised learning and supervised learning. We cannot use testing data to measure how the model performs. What we have are very general clusters. And if we even do the same algorithm again, it's possible to get different result. And we cannot even say which result is better. There are still some distinctions in unsupervised learning. If we want to cluster samples based their "similarity", we can use some methods like K-means and Latent Dirichlet Allocation. Those are methods used very often in natural language processing. If we don't want to cluster them, some other statistical methods like Principal Component Analysis can also be used.

About the data generating process, I think this is more related to what kind of data set we are working on. For some data sets, we have clear aim on what we want to study. Like if I am working on a bank customer churn project, whether the customer churn or not would be a great Y to start on. However, for some other data sets, like if we want to build a movie recommend system, we don't have a clear target to work on. Therefore, we could use some algorithms like KNN to cluster the similar users and recommend some movies they are most likely to watch.

Linear Regression

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## -- Conflicts ----------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

1. Using the *mtcars* dataset in R.

a.

```r
reg1 <- lm(mpg ~ cyl, data=mtcars)
summary(reg1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9814 -2.1185  0.2217  1.0717  7.5186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
## cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

The output is shown above. In this regression, β_0 is 37.8846 and β_1 is -2.8758.

b. The statistical form is $mpg_i = \beta_0 + \beta_1 * cyl_i + \epsilon_i$.

c.

```r
reg2 <- lm(mpg ~ cyl + wt, data=mtcars)
summary(reg2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

The result is shown above. In this case, $mpg = 39.6863 + (-1.5078)cyl + (-3.1910)wt$. The absolute value of coefficient of cylinder becomes smaller after adding weight variable, changing from -2.8758 to -1.5078. Therefore, it's fair to say that after adding weight, the effect of cylinder to MPG becomes smaller. Also, the coefficient of weight is -3.1910 and p-value is pretty small, suggesting that weight and cylinder both have signficant influence on MPG. We also note that both coefficients are negative, suggesting that cars with more cylinders and more weight tend to be less efficient.

d.

```r
reg3 <- lm(mpg ~ cyl + wt + cyl*wt, data=mtcars)
summary(reg3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + cyl * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2288 -1.3495 -0.5042  1.4647  5.2344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.3068     6.1275   8.863 1.29e-09 ***
## cyl          -3.8032     1.0050  -3.784 0.000747 ***
## wt           -8.6556     2.3201  -3.731 0.000861 ***
## cyl:wt        0.8084     0.3273   2.470 0.019882 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 28 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8457
## F-statistic: 57.62 on 3 and 28 DF,  p-value: 4.231e-12
```

In this case, the regression is $mpg = 54.3068 + (-3.8032)cyl + (-8.6556)wt + 0.8084cyl * wt$. The R-squared is pretty the same, from 0.8185 to 0.8457. The signficance level of cyl and wt are pretty the same with before. The coefficient of cyl and wt are changing, from -1.5078 to -3.8032 and from -3.1910 to -8.6556 respcetively. The interaction term is positive. By including the interaction term, we are theoretically asserting that there is non-linear relationship which somewhat correlates to both cyl and wt. That is to say, changing weight or changing cylinder are not independent on the effects of MPG, but rather the effects of changing in one variable depends on the other variable.

Non-linear Regression

1. Using the wage_data file.

a.

```
wage <- read_csv('wage_data.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   year = col_double(),
##   age = col_double(),
##   maritl = col_character(),
##   race = col_character(),
##   education = col_character(),
##   region = col_character(),
##   jobclass = col_character(),
##   health = col_character(),
##   health_ins = col_character(),
##   logwage = col_double(),
##   wage = col_double()
## )
```

```
w1 <- lm(wage ~ age, data=wage)
summary(w1)
```

```
##
## Call:
## lm(formula = wage ~ age, data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.265  -25.115   -6.063   16.601  205.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.70474    2.84624   28.71   <2e-16 ***
## age          0.70728    0.06475   10.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 40.93 on 2998 degrees of freedom
## Multiple R-squared:  0.03827,    Adjusted R-squared:  0.03795
## F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

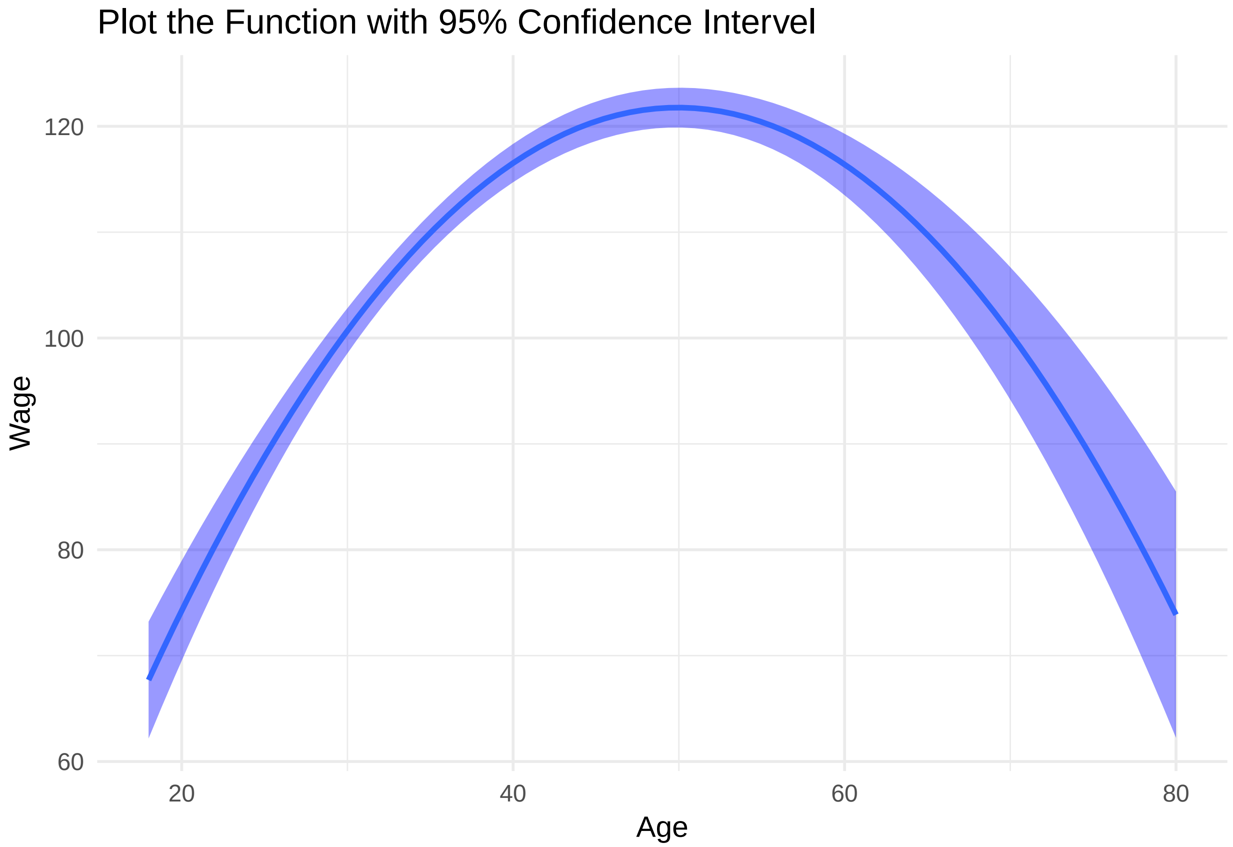
```r
w2 <- lm(wage ~ age + I(age^2), data=wage)
summary(w2)
```

```
##
## Call:
## lm(formula = wage ~ age + I(age^2), data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.425224   8.189780  -1.273    0.203
## age           5.294030   0.388689  13.620   <2e-16 ***
## I(age^2)     -0.053005   0.004432 -11.960   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2997 degrees of freedom
## Multiple R-squared:  0.08209,    Adjusted R-squared:  0.08147
## F-statistic:   134 on 2 and 2997 DF,  p-value: < 2.2e-16
```

The result is shown as above. The regression reports as $wage = -10.4252 + 5.2940age + (-0.0530)age^2$. Comparing with regression 1 (without polynomial term), the coefficient of age and age-squared are both pretty significant, suggesting that the relationship between age and wage is not linear. The coefficient of age is positive and age-squared is negative, suggesting that as age grows, the wage will increase at first and then decrease. For both regressions, the R-squared is pretty small, only like 8%. This is not superising because it is commonly known that a lot of factors contributing to the wage, and age could only explains part of it.

b.

```r
wage %>%
  ggplot(aes(x=age, y=wage), color = 'blue') +
  geom_smooth(method = lm, formula = y~poly(x, 2), fill='blue') +
  theme_minimal() +
  labs(x = 'Age',
       y = 'Wage',
       title = 'Plot the Function with 95% Confidence Intervel')
```



c. The plotting output is like a parabola. As age increases, the wage also increases early in young-age people's careers and then decreases after the peak. The peak is around the age of 50 and then the line declining. Additionally, we see that the confidence intervel widens out in the very beginning of people's career and also in late life. This makes sense because people start their career in different age and usually climbs to the peak in their mid-age. After that, people choose whether to retire or not based on different reasons, and even very few of them will contiune their career till the end of their life.

d. The linear regression depicts linear relationship between X and Y, where polynomial regression has the ability to depict more complex relationship between X and Y. Therefore, the polynomial regression has more flexibility to fit the data. Statistically, it is much easier for us to interpret the coefficient and confidence intervel in linear regression. On the other hand, polynomial regression is more powerful in fitting the data set and preditictions.