PS1

bohan yin

1/15/2020

Statistical and Machine Learning (25 points)

1. Describe in 500-800 words the difference between supervised and unsupervised learning.

- Supervised learning and unsupervised learning are two different kinds of machine learning techniques. They are different in several aspects. In supervised learning, we train the machine using data that is supervised. That is, the data is well-labelled so that you know what each part of data represents. So it can be interpreted as having a teacher to supervise you on understanding concepts. In terms of the relationship between the X's and Y, supervised learning learns the model where we know the algorithm to map X and Y (Y=f(X)). The target we are interested in is to optimize the estimation of the mapping function (Y=f(X)) so that we can use that to predict the output of future observation, or new X.

- Supervised learning can be analyzed into two groups, regression and classification. A regression problem is when the output variable is quantitative, or contiunous, which can be real value, such as heights, income and etc. For example, if we want to explore the effect of different education levels on personal income from a given dataset, we can use regression method to achieve that goal. A classification problem is when the output variable is qualitative, or categorical, which can be "pregnant" or "not pregnant" or etc. For example, if we want to determine whether the applicants of scholarship are from poor family or not, we can use classification method to achieve that.

- In unsupervised learning, the data that the machine uses is usually unlabelled, meaning that only input data will be given, so we need to let the model to discover hidden patterns and information of the data. In terms of the relationship between the X's and Y, unsupervised learning does not know the algorithm between existing data X and no corresponding output Y in there. The target we are interested in is to find underlying patterns in data, and find features in data for categorization. Since it is unsupervised, there is no teacher to guide us to understand more information in data.

- Unsupervised learning can be analyzed into two groups as well, clustering and association. A clustering type is when in unsupervised data, you want to find out a structure or feature in a collection of those data. For example, if we have a group of data that contains many flowers, we can use clustering to group those flowers in to different clusters, so that we can find certain patterns or species of those flowers. An association type is when you want to estalish associations among different variables in datasets. For example, we can use association method to discover the potential relationship between middle-aged people and the behavior of buying vitamin supplements, that people in middle-age are increasingly aware of their personal health, so they are more likely to purchase health products.

Linear Regression Regression (35 points)

1. Using the mtcars dataset in R (e.g., run names(mtcars)), answer the following questions:

a. Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

```
model1 <-  lm(mpg ~ cyl, mtcars)
tidy(model1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    37.9      2.07      18.3  8.37e-18
## 2 cyl            -2.88     0.322     -8.92 6.11e-10
```

The output for model1 is the miles per gallon (mpg), and the parameter values, which are beta0 and beta1, are 37.8846 and -2.8758 respectively. The first number means that in this model, keeping all parameters 0, the expected value of mpg is 37.8846. The second number means that in this model, every one unit increase of cylinders causes 2.8758 decrease in expected car's miles per gallon.

b. Write the statistical form of the simple model in the previous question (i.e., what is the population regression function?).

Y(mpg) = 37.8846 - 2.8758cyl

c. Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

```
model2 <- lm(mpg ~ cyl + wt, mtcars)
tidy(model2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    39.7      1.71      23.1  3.04e-20
## 2 cyl            -1.51     0.415     -3.64 1.06e- 3
## 3 wt             -3.19     0.757     -4.22 2.22e- 4
```

The intercept changes from 37.8846 to 39.6863, the coefficient of cylinders changed from -2.8758 to -1.5078, and the newly added coefficient of vehicle weight is -3.1910. We can interpret them as: in model2, keeping all parameters 0, the expected value of mpg is 39.6863; in model2, every one unit increase of cylinders causes 1.5078 decrease in expected car's mpg; every one unit increase of car's weight causes 3.1910 decrease in expected car's mpg.

d. Interact weight and cylinders and report the results. What is the same or different? What are we theoretically asserting by including a multiplicative interaction term in the function?

```
model3 <- lm(mpg ~ cyl + wt + cyl*wt, mtcars)
tidy(model3)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   54.3       6.13       8.86 0.00000000129
## 2 cyl           -3.80      1.01      -3.78 0.000747
## 3 wt            -8.66      2.32      -3.73 0.000861
## 4 cyl:wt         0.808     0.327      2.47 0.0199
```

The parameter values in model3 are different from model2 and model1, and they all increase in terms of absolute value. The interaction term here (cyl*mt) means that we assume that the effect of cylinders on mpg is dependent on the change of car weight. However, in model3, the p value for the interaction term implies that this interaction effect is not statistically significant.

Non-linear Regression (40 points)

1. Using the wage_data file, answer the following questions:

a. Fit a polynomial regression, predicting wage as a function of a second order polynomial for age. Report the results and discuss the output (hint: there are many ways to fit polynomials in R, e.g., I, ^, poly(), etc.).

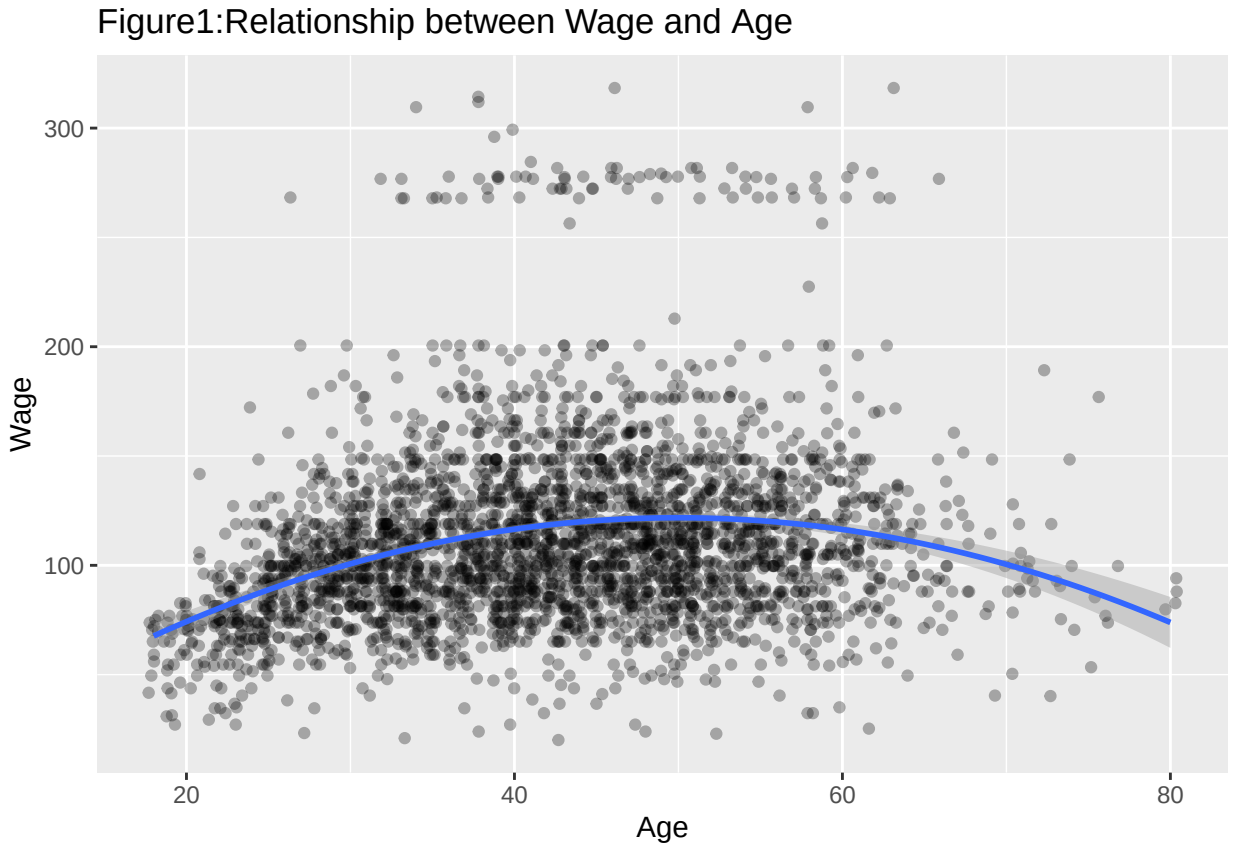
```
wage_data <- read.csv('wage_data.csv')
wgmodel <- lm(wage ~ age + I(age^2), wage_data)
tidy(wgmodel)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -10.4      8.19        -1.27 2.03e- 1
## 2 age           5.29     0.389       13.6  4.93e-41
## 3 I(age^2)     -0.0530   0.00443    -12.0  3.08e-32
```

In this model, the intercept (beta0) is -10,42522426, meaning that keeping all X being 0, the expected value for wage is -10,42522426. The coefficient for age is 5.29403003, meaning that each one unit increase in age increases expected wage by 5.29403003. The quadratic term age^2 means that the effect of age changes when people gets old. When you have 0 years of tenure, the slope is such that the wage would increase by 5.29403003 for one unit increase of age if the slope would remain unchanged, which it doesn't. Each additional age increase reduces the slope by 0.05300507. In this case the coefficient of the square term is negative, so the relationship is concave.

b. Plot the function with 95% confidence interval bounds.

```
wage_data %>% ggplot(aes(x = age, y = wage)) +
   geom_jitter(alpha = 0.3) +
   geom_smooth(se = TRUE, level = 0.95, method = 'lm', formula = y~poly(x,2)) +
   labs(x = 'Age',
        y = 'Wage',
        title = 'Figure1:Relationship between Wage and Age')
```



c. Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression?

In Figure1, we can see that, regression line is concave, and it shows a relatively quadratic line and a quadratic relationship. We can see that as age grows, wage goes higher, but when people reaches their 50s, wage stops to increase and starts to decrease. This is reasonable, as people getting old before a certain age, their working experience and competence for work is accumulating. When they reach a certain age, the aging reduces their productivity and health condition, thus lowering their income. Polynomial regression matches the data more accurately compared with no polynomial term added.

d. How does a polynomial regression differ both statistically and substantively from a linear regression (feel free to also generalize to discuss broad differences between non-linear and linear regression)?

Linear regression measures the linear correlations among variables, and polynomial regression does a better job of analysis when the relationship does not look linear at all. The polynomial regression is a linear in parameters, but the relationship beneath that is not linear. For example, in Figure 1, we can see that wage and age presents a quadratic relationship, so polynomial regression presents more accurate result than linear regression in this case. Statistically, polynomial regression has higher accuracy when the relationship is not linear; substantively, it has higher predicting power.