Problem Set 1: Learning and Regression

Xinyu LIU

2020/1/14

1. Statistical and Machine Learning

In general, any machine learning problem can be assigned to one of two broad classifications: *supervised learning* and *unsupervised learning*.

Supervised Learning

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Supervised learning problems are categorized into *"regression"* and *"classification"* problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some *continuous function*. In a classification problem, we are instead trying to predict results in a *discrete output*. In other words, we are trying to map input variables into discrete categories.

To describe the supervised learning problem slightly more formally, the goal is, given a training set, to learn a function h : X $\rightarrow$ Y so that h(x) is a *"good" predictor* for the corresponding value of y. In particular there will be proper cost function to measure the quality of the training. The goal is to find a set of parameters that minimize the cost function of the training data.

Unsupervised Learning

While supervised learning algorithms need labeled examples (x,y), unsupervised learning algorithms need only the input (x). It allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables. We can derive this structure by *clustering* the data based on relationships among the variables in the data. With unsupervised learning there is no feedback based on the prediction results.

Two of the main methods used in unsupervised learning are *principal component* and *cluster analysis*. Cluster analysis is used in unsupervised learning to group, or segment, datasets that has not been labelled, classified or categorized with shared attributes in order to extrapolate algorithmic relationships.Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

Another frequently used technique based on unsupervised learning is dimension reduction, which helps to reduce the dimension of problem and captures the most important characteristics.

2. Linear Regression

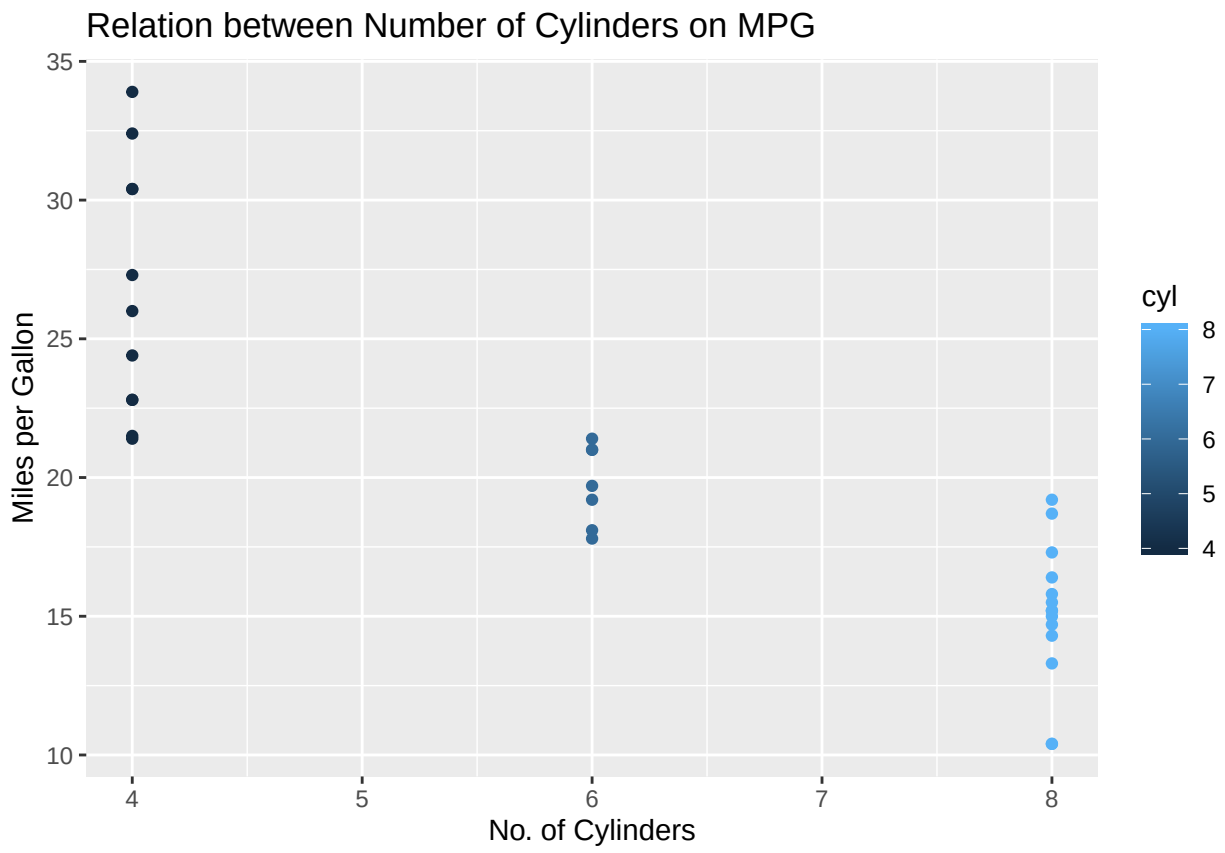
```
# mtcars: Motor Trend Car Road Tests
# The data was extracted from the 1974 Motor Trend US magazine,
# and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles
# 1. Loading
data("mtcars")
# 2. Print
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a. Predict miles per gallon (mpg) as a function of cylinders (cyl)

The output and parameter values for my model are shown in the summary bellow.

```
library(ggplot2)
#Doing a quick scatter plot yields the following
qplot(cyl, mpg, data = mtcars, colour = cyl, geom = "point",
      ylab = "Miles per Gallon", xlab = "No. of Cylinders",
      main = "Relation between Number of Cylinders on MPG")
```



```r
model1 <- lm(mpg ~ cyl, data = mtcars)
summary(model1)
```

```
##
## 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
```

b. Write the statistical form of the simple model

The linear model equation can be written as follow:

$$mpg = 37.8846 - 2.8758 * cyl$$

c. Add vehicle weight (wt) to the specification The results are shown in the summary bellow.

```r
model2 <- lm(mpg ~ cyl + wt, data = mtcars)
summary(model2)
```

```
##
## 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
```

Discussion:

Frist of all, the coefficient of cylinder in both regression seem to be statistically significant as negative numbers. After adding weight into "model1", its coefficient's absolute amount decreased a bit in the mutilinear regression "model2", meaning the negative influence of weight on miles per gallon can partially explain (or absorb) the negative effect of number of cylinders on that. Since both of the coefficient are significant and adjusted R^2 increased from 72% to 82%, I conclude that model2 better captures the characteristics of "mpg" in that it improves R^2.

d. Interact weight and cylinders

The results are shown in the summary bellow.

```
model3 <- lm(mpg ~ cyl + wt + cyl*wt, data = mtcars)
summary(model3)
```

```
##
## 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
```

Discussion:

The direction of "cyl" and "mpg" stays the same (negative), but the absolute magnitude increases. The theoretical assertion is that the effect of increase of "cyl" on "mpg" depends on the amount of "wt".

3.Non-linear Regression

a. fit a polynomial regression, predicting wage as a function of a second order polynomial for age

```
wage_data<-read.csv('wage_data.csv',header=TRUE)
head(wage_data)
```

```
##        X year age           maritl     race       education
## 1 231655 2006  18 1. Never Married 1. White    1. < HS Grad
## 2  86582 2004  24 1. Never Married 1. White 4. College Grad
## 3 161300 2003  45       2. Married 1. White 3. Some College
## 4 155159 2003  43       2. Married 3. Asian 4. College Grad
## 5  11443 2005  50      4. Divorced 1. White      2. HS Grad
## 6 376662 2008  54       2. Married 1. White 4. College Grad
##                region       jobclass         health health_ins  logwage
## 1 2. Middle Atlantic  1. Industrial      1. <=Good      2. No 4.318063
## 2 2. Middle Atlantic 2. Information 2. >=Very Good      2. No 4.255273
## 3 2. Middle Atlantic  1. Industrial      1. <=Good     1. Yes 4.875061
## 4 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes 5.041393
## 5 2. Middle Atlantic 2. Information      1. <=Good     1. Yes 4.318063
## 6 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes 4.845098
##        wage
## 1  75.04315
## 2  70.47602
## 3 130.98218
## 4 154.68529
## 5  75.04315
## 6 127.11574
```

```
model4 <- lm(wage ~ poly(age, 2, raw=TRUE), data = wage_data)
summary(model4)
```

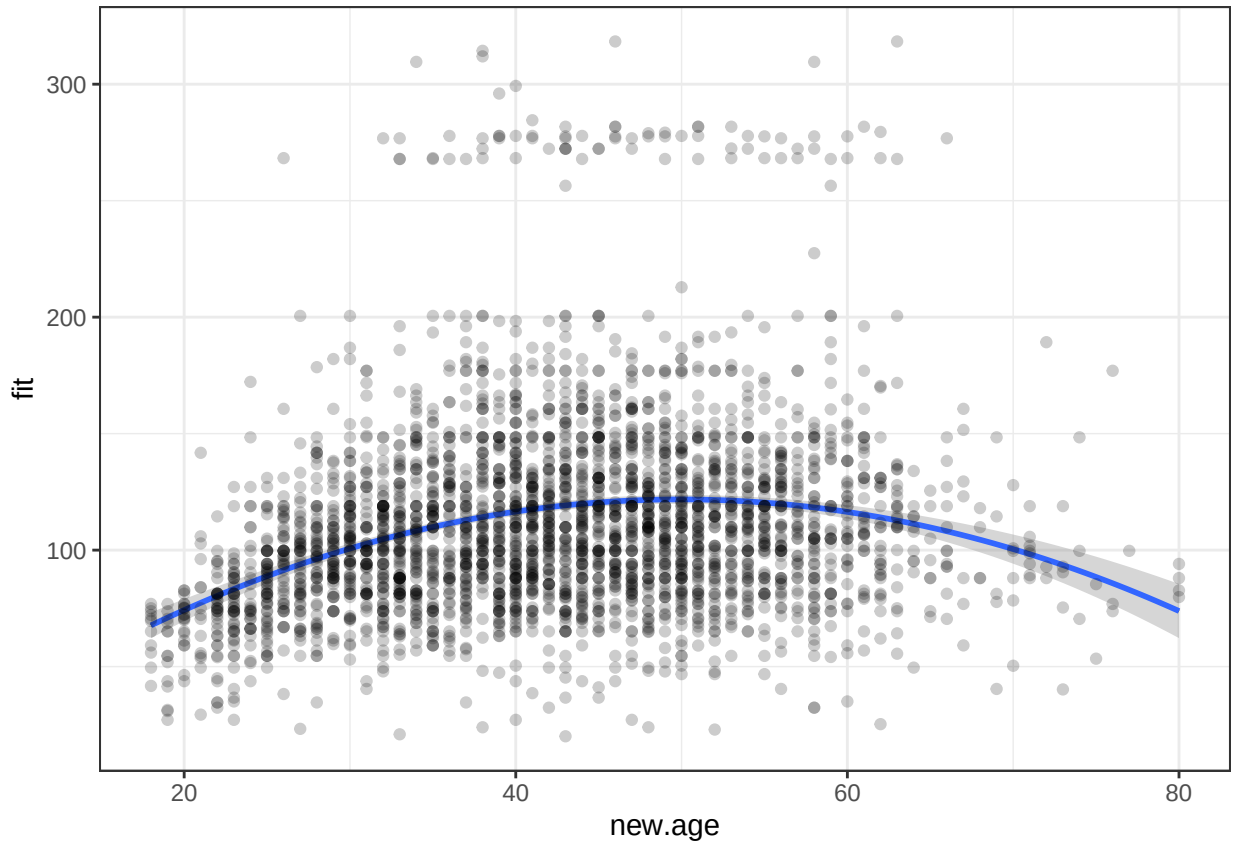
```
##
## Call:
## lm(formula = wage ~ poly(age, 2, raw = TRUE), data = wage_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -10.425224   8.189780  -1.273    0.203
## poly(age, 2, raw = TRUE)1   5.294030   0.388689  13.620   <2e-16 ***
## poly(age, 2, raw = TRUE)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
```

b. plot the function with 95% confidence interval bounds

```
new.age <- seq(min(wage_data$age), max(wage_data$age), length.out=100)
age <- data.frame(age=new.age)
err <- predict(model4, newdata = age, se.fit = TRUE)

age$lci <- err$fit - 1.96 * err$se.fit
age$fit <- err$fit
age$uci <- err$fit + 1.96 * err$se.fit

ggplot(age, aes(x = new.age, y = fit)) +
  theme_bw() +
  geom_line() +
  geom_smooth(aes(ymin = lci, ymax = uci), stat = "identity") +
  geom_point(data = wage_data, aes(x = age, y = wage), alpha = 0.2)
```



c. describe the output

Both the 1st order and 2nd order are statistically significant. In general wage is a concave function of age, with the peak appearing in the middle rage of age around 50. By using this polynominal model we assert that the relationship between wage and age is polynominal.

d. How does a polynomial regression differ both statistically and substantively from a linear regression?

Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y|x)$ is linear in the unknown parameters that are

estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression. One big difference between linear regression and polynomial regression is that regressors in polynomial regression can be dependent with each other, for instance x and x^2.

As for potential bias between linear and polynomial regression, due to the simplicity of linear regression, it tends to underfit the model therefore pocesses relatively higher bias. Whileas polynomial regression is able to capture more complicated feature but may also have higher variance.