Gill_Sarah_ML_PS_1

Sarah Gill

1/13/2020

Problem Set 1: Learning and Regression

Introduction to Machine Learning

Statistical and Machine Learning

Supervised learning is a means of producing predictive models given data where the features and labels are known. The learner is able to improve (reduce the error), over iterations by responding to feedback. Learning is conceptualized as the machine's response to this feedback (it's error rate). Thus the data, with both X and Y, is used to iteratively refine the algorithm or mapping process of X to Y as the machine tries to minimize the error. This is much like regression analysis if we were looking only for the best model of the data regardless of theory. The target of supervised learning is to fit a model well to the dataset (given X and Y) such that it accurately predicts the Y for data outside of the dataset. Initially this means doing well in the test portion of the data (kept out from the initial training dataset), but in the end the idea is to be able predict Ys that are not known. For instance, supervised machine learning is often used in financial forecasting.

Unsupervised learning is a means of discovering patterns in data, either directly or through dimension-reduction. With unsupervised learning the dataset only has the features (X), and not the labels (Y). The learning is not supervised in that the machine is not given the feedback of how well it did in each iteration, as is the case in supervised learning, but instead engages in pattern discovery. The number of classes does not need to be known. The goal is not to create a predictive model, but rather to learn about the data. For example, unsupervised learning has been used to generate possible protean folding combinations for researchers to investigate in a bio-chemistry research lab.

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
```

```
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

1. Linear Regression Regression

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

```
# Here we regress wage on IQ and then on age, bivariate
#lm(wage ~ IQ, data = wage)

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
```

```
#predict(model1)
```

There is a statistically significant relationship between number of cylinders and mpg. Each additional cylinder is associated with a decrease in miles per gallon of 2.876mpg. These results show a theoretical mpg of 37.88mpg for the impossible car that has no cylinders.

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

mpg = intercept + b*cyl + error

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, 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
```

```
#predict(model2)
```

There is a statistically significant relationship between number of cylinders and mpg and between weight and mpg. Each additional cylinder is associated with 1.508 fewer miles per gallon, holding weight constant. Each additional 1000lbs (one unit in wt) is associated with 3.191 fewer miles per gallon, holding number of cylinders constant.

These results show a theoretical mpg of 39.69mpg for the impossible car that weighs nothing and has no cylinders.

Note that the magnitude of the estimate for cyl is less than in the univariate regression. Some of the decrease in mpg that was associated with cly in the previous model may be better explained by weight.

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, data = mtcars)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ 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
```

```
#predict(model3)
```

When we include an interaction term for wt and cyl, the coefficient estimates for both wt and cyl increase in magnitude (they are both more negative). However these coefficients can no longer be interpreted on their own. By including an interaction term we are making the assertion that the relationship between weight and number of cylinders moderates or mediates the relationship between weight and mpg and between number of cylinders and mpg. In other words, we assert that the relationship between cylinders and mpg is effected by weight and the relationship between weight and mpg is effected by number of cylinders. This assertion is supported by the results, with the estimate for the interaction term being statistically significant, and it's inclusion improving our Adjusted R-squared. The estimated relationship is:

cyl on mpg: -3.81 + 0.81*wt wt on mpg: -8.66 + 0.81*cyl

So for a car that is the mean weight in this dataset, an increase of one additional cylinder is associated with a decrease of roughly 1.20mpg.

mean wt in dataset: 3.217

```
 -3.8032 + 0.8084 *(3.217)
```

```
## [1] -1.202577
```

Or for a car that has the mean number of cylinders for this dataset, an increase of an additional 1000lbs (one unit of wt) is associated with a decrease of roughly 3.65mpg

mean cyl in dataset: 6.188

```
-8.6556  + 0.8084 *(6.188)
```

```
## [1] -3.653221
```

2. Non-linear Regression

a. Fit a polynomial regression, predicting wage as a function of a second order polynomial for age. Report the results and discuss the output

```r
wage_data <- read_csv("wage_data.csv")
```

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

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   year = col_integer(),
##   age = col_integer(),
##   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()
## )
```

```r
wage_model <- lm(wage ~ age + I(age^2), data = wage_data)
#wage_model <- lm(wage ~ poly(age, degree = 2, raw = T), data = wage_data)
summary(wage_model)
```

```
##
## Call:
## lm(formula = wage ~ age + I(age^2), 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
## 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
```

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

By fitting a polynomial to this data, we are asserting that the relationship between age and wage may be different at different ages. Using a quadratic function we are allowing for the relationship to be concave or convex and even to change signs.

Here we see that at lower ages increasing age is associated with increasing wage, however this is increasing at a decreasing rate (as indicated by the negative sign on the age^2 coefficient, thus our fitted curve is concave). Also the positive slope may change to negative at some age.

increasing age by one year is associated with a change in wage of 5.29 - 0.053(age)

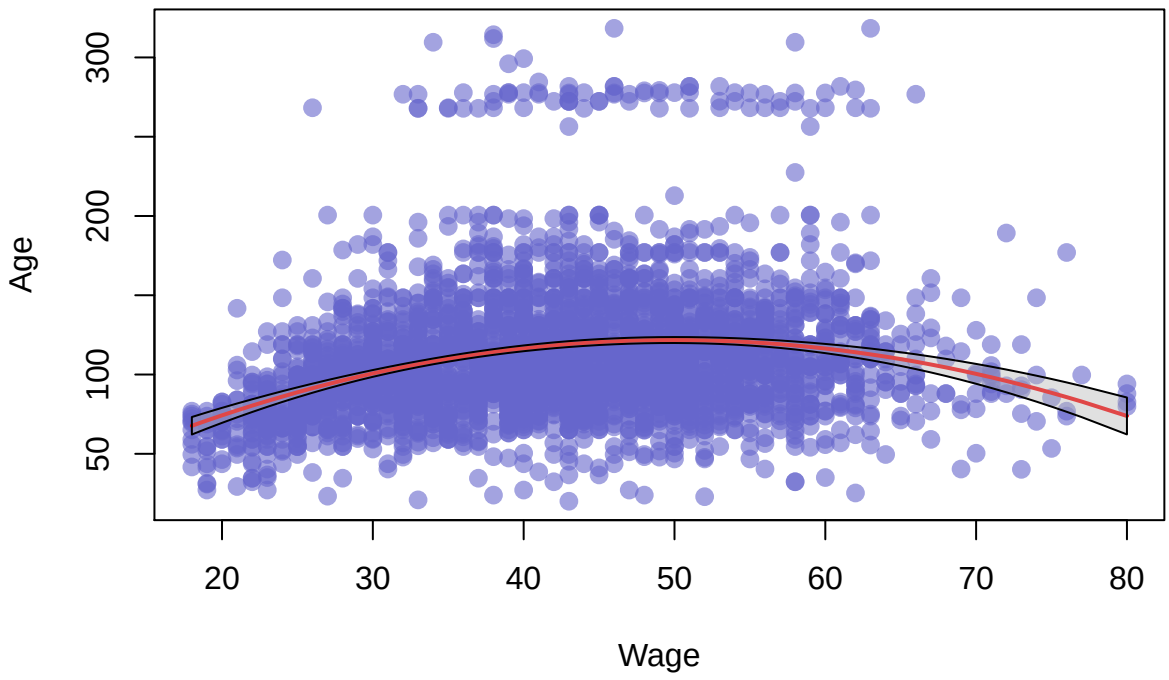
b. Plot the function with 95% confidence interval bounds.

```
# plot
#works but throws an error
y = wage_data$wage
x = wage_data$age
plot(x,y,col=rgb(0.4,0.4,0.8,0.6),pch=16 , cex=1.3, xlab = "Wage", ylab = "Age")

#myPredict <- predict( wage_model )
#ix <- sort(x,index.return=T)$ix
#lines(x[ix], myPredict[ix], col=2, lwd=2 )

#Curve
myPredict <- predict( wage_model , interval="predict")
ix <- sort(x,index.return=T)$ix
lines(x[ix], myPredict[ix , 1], col=2, lwd=2 )

CI <- predict(wage_model, x=wage, interval = 'confidence', level=0.95)
#ix <- sort(x,index.return=T)$ix
#lines(x[ix], CI[ix , 1], col=2, lwd=2 )
#CI
polygon(c(rev(x[ix]), x[ix]), c(rev(CI[ ix,3]), CI[ ix,2]), col = rgb(0.7,0.7,0.7,0.4) , border = "blac
```




```
#cite: https://www.r-graph-gallery.com/44-polynomial-curve-fitting.html
#cite: https://www.r-graph-gallery.com/45-confidence-interval-around-polynomial-curve-fitting.html
#cite: https://www.researchgate.net/post/How_can_I_put_confidence_intervals_in_R_plot
```

From the plot we can see that wage increases with age until about 50, when each additional year of age is then associated with a decrease in wage.

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

```
lin_wage_model = lm(wage ~ age, data = wage_data)

anova(lin_wage_model, wage_model)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ age + I(age^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.04 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#cite: https://www.youtube.com/watch?v=ZYNOYD7UfK4
```

The polynomial regression allows for a non-linear relationship between age and wage and thus is a more flexibal though harder to interpret modle. The model that includes age squared is statistically significantly better at predicting age than the purely linear model.