problem-set-1

See syllabus for submission details.

Statistical and Machine Learning (25 points)

1. Describe in 500-800 words the difference between supervised and unsupervised learning. As you respond, consider the following few questions to guide your thinking, e.g.: • What is the relationship between the X's and Y? • What is the target we are interested in? • How do we think about data generating processes? • What are our goals in approaching data? • How is learning conceptualized? And so on. . .

Supervised machine learning uses a set of features X from a population n to predict some categorical or continuous outcome Y. The Y is a known quantity or quality, and therefore the accuracy of the model can be ascertained by comparing the predicted and actual Y. The target is to predict this quantity Y in a different testing data set. The goal in approaching this data is to learn how these X features interact with each other and the outcome Y.

Unsupervised machine learning uses a set of features X from a population n to predict some latent characteristic Y. This latent characteristic is usually group membership into 1 of k groups. The relationship between X and Y is more complex as it is not explicitly evident in the data. The assumption is that the population n is made up of more homogeneous subgroups K, and the X features can be used to discover which of the K groups an individual belongs to. The goals is to find underlying substructure to the data that is assumed to be a pooling of several clusters or subgroups.

Linear Regression Regression (35 points)

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

a. (10) Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

```
data(mtcars)
head(mtcars)
linearModel <- lm(mpg ~ cyl, data=mtcars)
print(linearModel)
summary(linearModel)
```

```
 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

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

mpg[ij] = B0 + B1*(cyl[ij]) + E[ij]

c. (10) Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

```
linearModel <- lm(mpg ~ cyl+ wt, data=mtcars)
print(linearModel)
```

```
summary(linearModel)

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
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
```

Cylinder is still inversely associated with miles per gallon, however the coefficient point estimate is smaller. Additionally, the standard error of hte coefficient is slightly larger. Weight is also inversely associated with miles per gallon.

d. (10) 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?

```
mtcars$interaction = mtcars$wt*mtcars$cyl
linearModel <- lm(mpg ~ cyl+ wt + interaction, data=mtcars)
print(linearModel)
summary(linearModel)

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
interaction   0.8084     0.3273   2.470 0.019882
```

The coefficient for both cylinder and weight is much larger, as are the standard errors for both. The interaction term is also significant. We are asserting that increases in both cylinder and weight have a multiplicative effect on the resulting miles per gallon; so in addition to the increase in miles per gallon from an increase in cylinder size and the increase in miles per gallon from an increase in weight, there is an increase from the interaction of these two variables that is greater than the addition of the two effects.

Non-linear Regression (40 points)

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

a. (10) 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.).

```
library(foreign)
wage_data <- read.csv("C:/Users/sbhavani/Desktop/wage_data.csv")
wage_data$age2 = wage_data$age^2
polyModel <- lm(wage ~ age + age2, data=wage_data)
print(polyModel)
summary(polyModel)

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
age2         -0.053005   0.004432 -11.960   <2e-16
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
```

Age is linearly related to wage, and each year increase in age results in $5.29 increase in age. Age is also quadratically related to wage, so that with each 1 unit increase in age squared, the wage decreases by $0.05.

b. (10) Plot the function with 95% confidence interval bounds.

```
library(ggplot2)
wage_data$prediction = predict(polyModel)

lm = predict(polyModel, interval = 'confidence')
lm = as.data.frame(lm)
lm$age = wage_data$age

png("wage_age.png")
ggplot(data = lm, aes(x = age, y = fit)) +
geom_smooth(data=lm,    aes(  ymin = lwr, ymax = upr ), stat="identity") +
theme(panel.background = element_rect(fill = "white")) +
ylab("Predicted wage")+
xlab("Age")
dev.off()
```

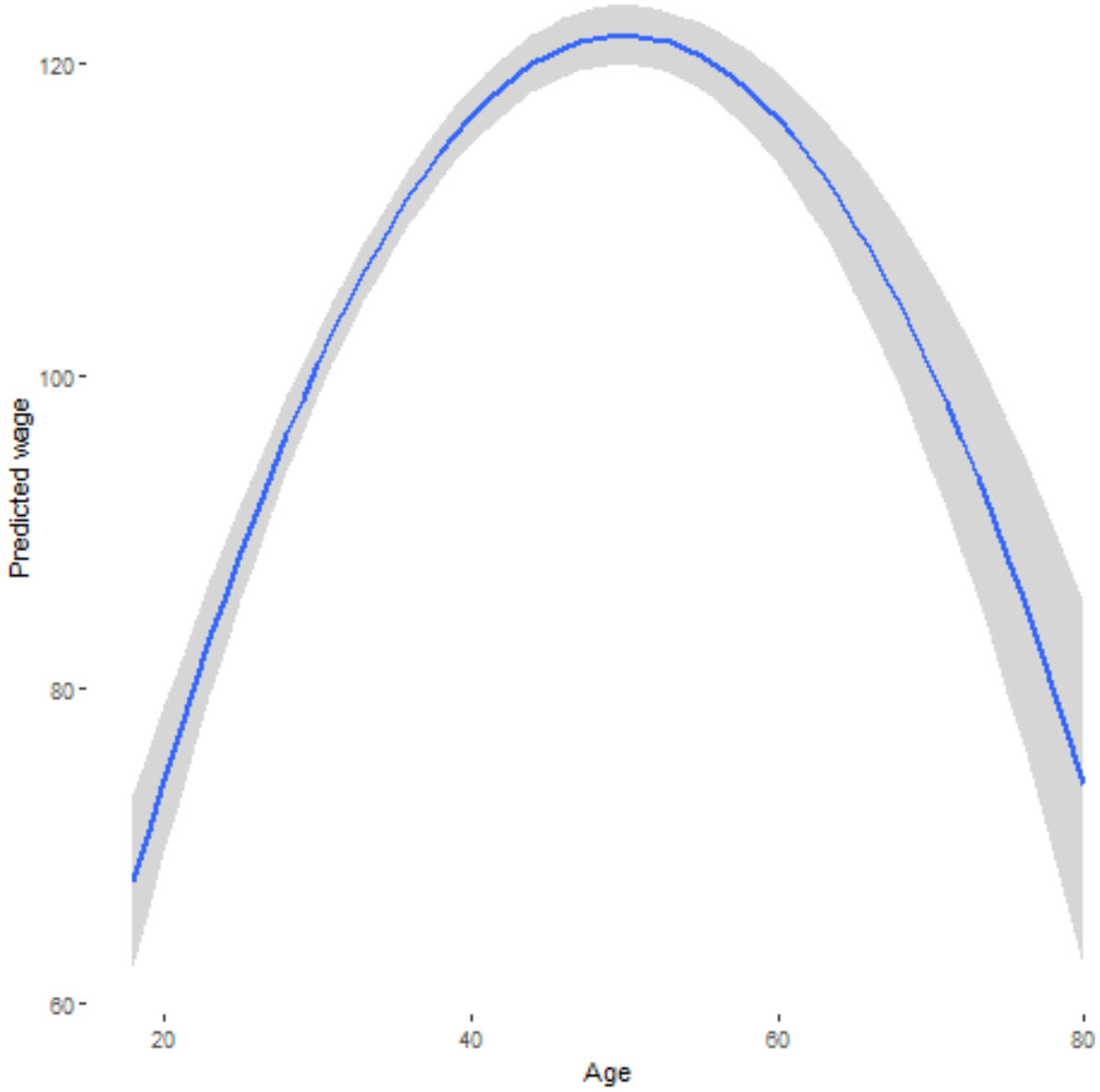

FIGURE 1

c. (10) Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression?

The confidence interval gets larger at the tail ends, and especially as age increases. The quadratic term asserts that there is a non-linear relationship between age and wage. Specifically, the negative coefficient for age-squared suggests that with increasing age, wage begins to fall.

d. (10) 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 assumes a strictly linear relationship between the feature X and outcome Y. With polynomial regression, the feature X is expanded into multiple features that are exponentiated results of X. In polynomial regression, the outcome Y is estimated using multivariable linear regression with the features being polynomial expansions of X. There is bound to be correlation between the linear, quadratic, cubic and other polynomial transformations of X. Further, with polynomial regression, extrapolation may be difficult as with increasing values of X, the higher order polynomial terms may have a greater impact on the predicted outcome Y.