Problem Set 1

Grace Wright

1/15/2020

Problem Set 1: Learning and Regression

Statistical and Machine Learning

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

There are two types of Machine Learning: supervised and unsupervised. Supervised is where you have defined inputs (X) as well as defined outputs (Y). The aim of the Machine is to make sense of the data, by giving the best approximation of the relationship between the inputs and outputs. This can come in the form of classification or regression. Classification, allows the data to be grouped or "classified" along common parameter lines. For example, all the oranges would be classified together as well as all the apples. Classification is primarily for categorical variables. On the contrary, regression is primarily used for quantitative data. Regression is also used in supervised machine learning, and similarly aims to find the best approximate relationship between the variables, but along a continuous y-value. Classification and regression are possible because the inputs and outputs are pre-defined, the machine is simply trying to make sense of it – the basic idea behind supervised machine learning. However, in an effort to create the perfect descriptive model, one must be wary of over-fitting the data. Overfitting the data is a common, but avoidable, error in supervised machine learning.

In unsupervised machine learning, the main goal is to look for a potential underlying structure in the data. Rather than telling the machine what code to follow, you teach it to observe patterns and define a potential underlying structure within which to understand the data. One of the most common tasks within unsupervised machine learning is clustering. Clustering is where the machine is trained to identify the pattern within the data and then clusters those similar elements together. For example, if given a large set of animal pictures, the machine could learn to cluster all the like animals together, as in all the horses in one group and all the dogs in another. Additionally, another method in unsupervised machine learning is dimensionality reduction. Dimensionality reduction allows one to learn the relationship between features and simplify into a few latent features. Due to its interest in simplifying broad features, dimensionality reduction is able to relatively quickly identify latent features.

As a broad view, within supervised machine learning, classification is used when working with discrete variables whereas regression is used when using continuous variables. Conversely, within unsupervised machine learning, clustering is used when working with discrete variables and dimensionality reduction is used when working with continuous variables. Supervised machine learning is helping you process through "labeled" data. A supervised machine helps you build and deploy a model that can aid you in predicting future outcomes. In juxtaposition to supervised machine learning, in unsupervised machine learning primarily works with "unlabeled" data and one allows the machine to uncover information within the data. One would choose to utilize unsupervised machine learning if they are aiming to find underlying patterns within the data, as in efforts to categorize of cluster. On the other hand, one would be inclined to use supervised machine learning if you want to build a model based on current existing "labeled" data in an effort to predict future outcomes, as in predicting how long a drive might take.

Linear Regression Regression

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

1a. Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

```
mtcars <- mtcars
? lm
mpgcyl_regress <- lm(data = mtcars , formula = mpg ~ cyl)
mpgcyl_regress
summary(mpgcyl_regress)
```

The coefficient for the intercept is 37.885, and for the cylinder is -2.876. The residuals for the min is -4.9814, for the first quartile is -2.1185, for the median is .2217, for the third quartile is 1.0717, and the max is 7.5186. The multiple R^2 is .7262 with a very small p-value.

1b. Write the statistical form of the simple model in the previous question (i.e., what is the population regression function?).

$Y_i = 37.8846 - 2.8758X_i$

1c. Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

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

The coefficient size for the interaction between mpg and the cylinder size, is different when you include the factor of weight. The mpg will change by factor of -1.5 for every one unit increase in the cylinder size. Likewise, the mpg will change by a factor of -3.19 for every unit change in the weight. This shows that the change in the mpg will be more greatly effected by a change in the weight than in a change in cylinder size. The change in the coefficient size of the cylinder once the weight is included shows that there is an interaction and covariance between the effect of weight on mpg and the effect of cylinder size on mpg. This shows that it is not only the cylinder size that effects the mpg, but also the weight, and even to a greater degree than the cylinder size.

1d. 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?

```
mpgclywt_regress2 <- lm(data = mtcars , formula = mpg ~ cyl*wt)
mpgclywt_regress2
summary(mpgclywt_regress2)
```

Theoretically we are asserting that there is an interaction between the cylinder size and weight as they effect the mpg. In other words, the combination of the cylinder size and weight effect the miles per gallon observed.

Non-linear Regression

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

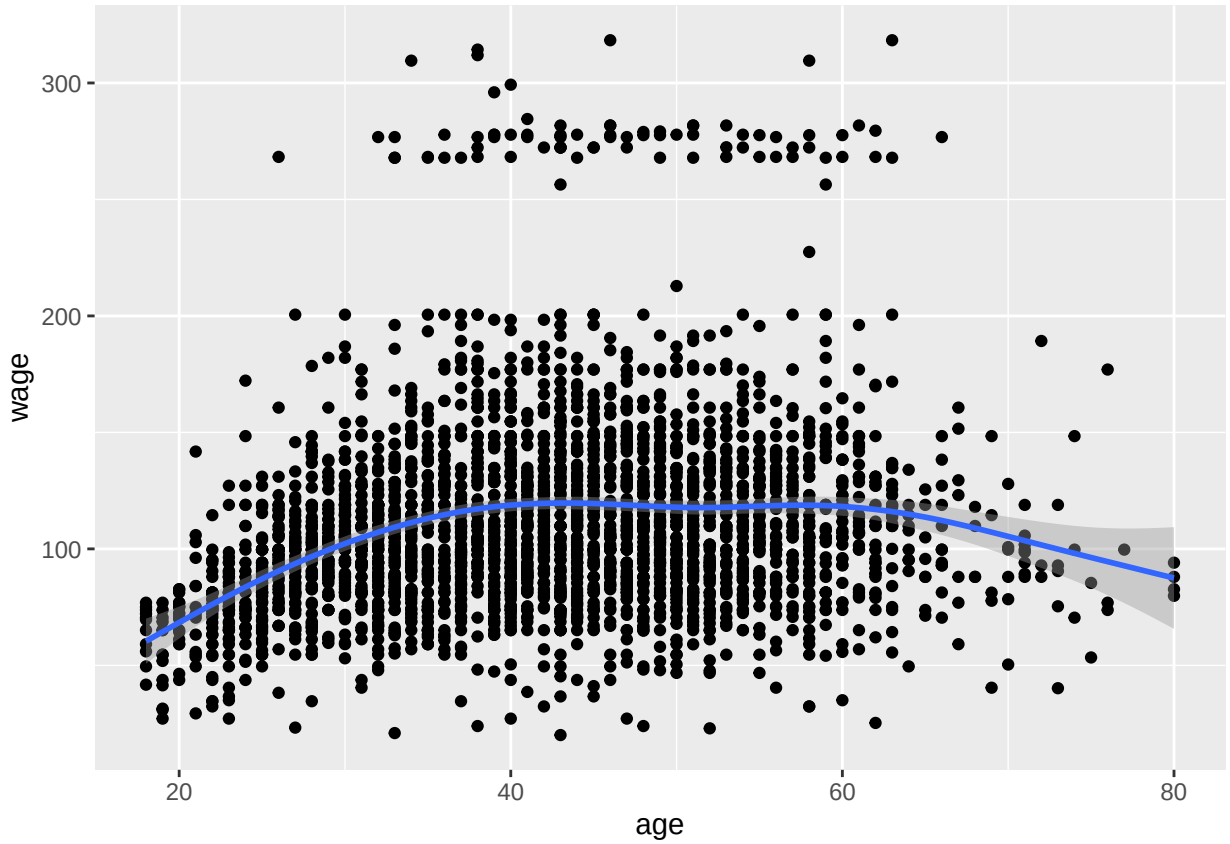
1a. Fit a polynomial regression, predicting wage as a function of a second order polynomial for age. Report the results and discuss the output (hint: there are many ways to fit polynomials in R, e.g., I, ^, poly(), etc.).

```
wage_data <- read.csv("wage_data.csv")
wage_data_regress <- lm(data = wage_data , formula = wage ~ age + I(age^2))
wage_data_regress
summary(wage_data_regress)
```

The outcome of the polynomial regression tells us that age and wage are correlated, where a younger individual is most likely to make the least income, with the average increasing in middle age, then starting to slope downwards for older individuals. The low R^2 value tells us that the values do not account for a large percentage of the variation between the variables. Additionally, the high residual standard error shows that this regression has little predictive power.

1b. Plot the function with 95% confidence interval bounds.

```
wage_data_plot <- ggplot(data = wage_data, aes(age,wage))+
  geom_point()+
  stat_smooth(formula = wage_data_regress$formula, level = .95)
wage_data_plot
```



1c. Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression? The regression line shows a concave linear relationship between an individual's age and wage. In other words, on average, an individuals wage starts out less when young, increases towards middle age, then decreases as individuals get older. By fitting a polynomial regression, we are asserting that the results are still linear (because the x value is squared rather than the Beta coefficient), but that they are best represented with a curve, or bend, rather than with a homogenously sloped line. Additionally, by fitting a polynonmial regression rather than a monomial regression we are asserting that age and wage do not change by a constant value, but rather by an exponential value. This polynomial regression is simply a more flexible linear regression.

1d. 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)? Substantively, both polynomial regression and linear regression can be classified as "linear", it simply depends upon the nature of their equations. If the equation has a common slope (b), even

if the X value is squared, the regression can be understood as linear. Statistically, as long as it follows the basic equation (y = mx + b), even with a longer chain or a squared x-value, it's linear. In linear regression, x is the explanatory variable and y is the dependent varibale. It becomes non-linear when there is not a common slope as evidenced in the common equation (y = mx + b). This polynomial regression is simply a more flexible linear regression. However, one must be wary of over-fitting.