Problem Set 1

Pete Cuppernull

1/17/2020

Statistical and Machine Learning

The two main types of machine learning tasks are supervised and unsupervised learning. In supervised learning, the researcher leverages a data set for which a certain selection of indicator variables are assumed to explain a particular outcome of interest. In this scenario, both the values of the indicators and the value of the outcome of interest are known. The researcher uses this data to "train" a learner to correctly predict the outcome of interest – if a satisfactory learner is built, it can then be used to identify the outcome of interest for observations where only the values of the indicators are known. Generally speaking, a researcher's goal in this process would be to create a learner with maximum predictive power, potentially saving valuable data processing resources as part of the process — a common way the judge the predictive power of the model would be to assess how accurately the model can predict values of Y on a holdout set of test data (a subset of the original data which was not used to train the model).

Alternatively, unsupervised learning tasks do not require the observations in the data to have a designated outcome of interest. In terms of X's and Y, this means that while supervised learners seek to use the values of both the X's and Y to predict unknown values of Y, unsupervised learners use data which do not have an explicit Y variable. Instead, unsupervised learning tasks seek to recover some latent structure of the data that might be unknown to the researcher – unsupervised learners create "buckets" of observations based on the provided data. Thus, unsupervised learners allow a researcher to learn more about the nature of that data at hand, for example by identifying clusters among certain characteristics of the observations. Because of the nature of the approach of unsupervised learning, researchers are generally not concerned with the "accuracy rate" of the learner, since there is no "correct" classification that the learner is trying to generate.

These two types of methods can be applied in a variety of computational settings. Supervised learning techniques are often leveraged to tackle regression problems. Here, the researcher assumes that a continuous outcome of interest can in some way be explained as a function of the available predictor variables. In the case of discrete outcomes, referred to as classification problems, supervised learners can still be applied – however, unsupervised learners can also be used to identify discrete characteristics in the data in the absence of an explicit outcome of interest.

It is important to note that both supervised and unsupervised learners prioritize the maximization of predictive power over understanding the data generating process. In statistical learning, alternatively, the goal of the researcher is often oriented towards understanding the data generating process. This distinction underlies why training and test data sets are heavily used in machine learning methods, and why full data sets are often used to construct models in statistical learning.

Linear Regression

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

```
linear_model <- lm(mpg ~ cyl, mtcars)

print(linear_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
```

```
##
## Coefficients:
## (Intercept)          cyl
##      37.885       -2.876
```

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

$$MPG = \beta_0 + CYL\beta_1 + \epsilon$$

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

```
linear_model2 <- lm(mpg ~ cyl + wt, mtcars)
print(linear_model2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl           wt
##      39.686       -1.508       -3.191
```

In adding weight to the specification, we can observe the value of the cylinder coefficient decrease from -2.876 to -1.508. From this, we can infer that some of the explanatory power that was attributed to the number of cylinders in a car in the first model was in fact actually due to the weight of the car. Both of the coefficients remain negative, though, suggesting that cars with more cylinders have lower gas mileage.

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?

```
linear_model3 <- lm(mpg ~ cyl + wt + cyl*wt, mtcars)
print(linear_model3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + cyl * wt, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl           wt       cyl:wt
##     54.3068      -3.8032      -8.6556       0.8084
```

We can observe that the intercept of the model increases considerably, up to 54.307, and the coefficients for both cylinders and weight significantly increase (they both remain negative, though, as in the other models). By including the multiplicative interaction term, we are asserting that the effect of at least one of the independent variables depends on the value of another independent variable – for example, we might believe that the effect of the weight of the car on MPG varies between cars with different numbers of cylinders.

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.

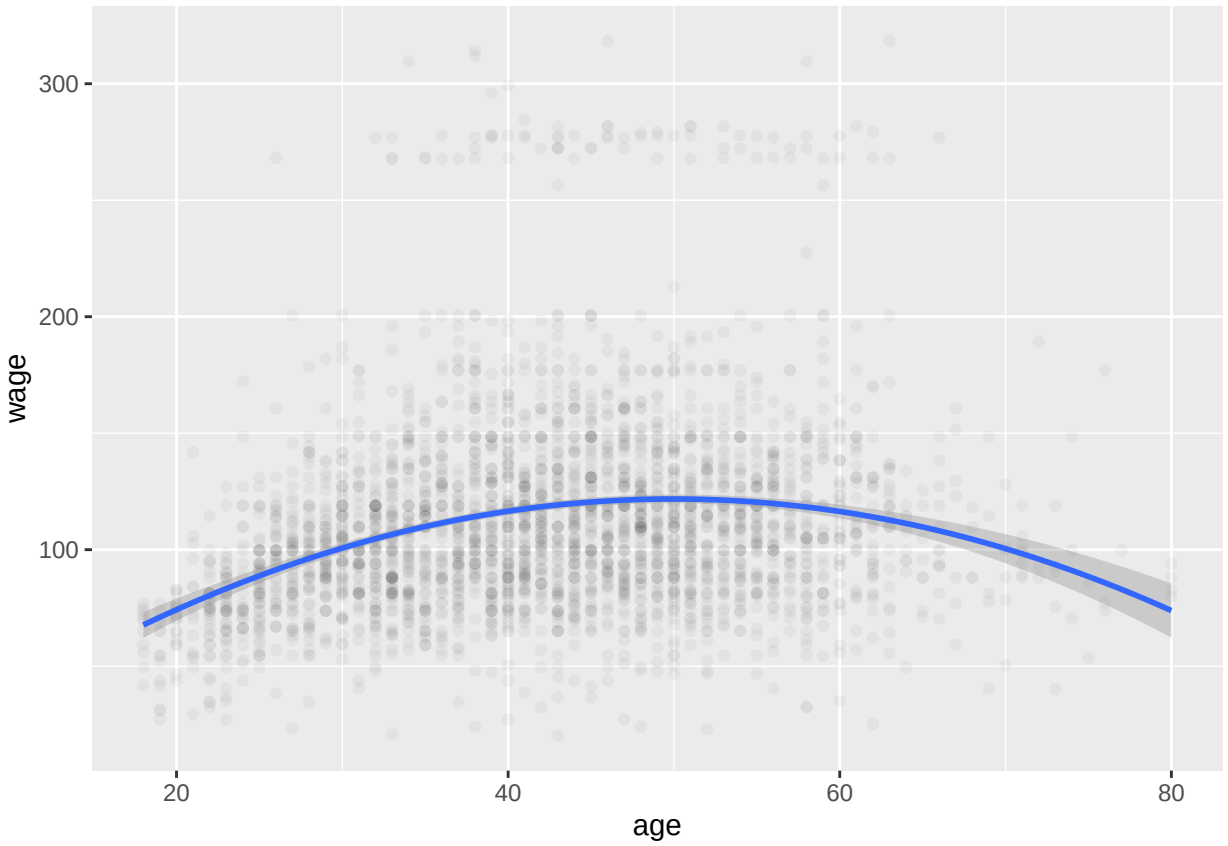
```
library(tidyverse)
wage_data <- read_csv("/Users/petecuppernull/Dropbox/UChicago/2019-20/Winter/Machine Learning/Problem Se

model_poly2 <- glm(wage ~ age + I(age^2),
                   data = wage_data)

print(model_poly2)
```

```
##
## Call:  glm(formula = wage ~ age + I(age^2), data = wage_data)
##
## Coefficients:
## (Intercept)          age     I(age^2)
##   -10.42522      5.29403     -0.05301
##
## Degrees of Freedom: 2999 Total (i.e. Null);  2997 Residual
## Null Deviance:       5222000
## Residual Deviance: 4793000   AIC: 30650
```

We can see that wage is heavily determined by age – the negative coefficient on the Age^2 term indicates that once age reaches a certain threshold, the negative effect of the Age^2 term will be larger than the positive effect of Age, meaning that the overall expected wage will begin to decrease at that threshold. Finally, the results display a shortcoming for this type of model selection – for some values of X (2 or below, 98 and above), negative point estimates are generated for wage, which are non-intuitive.

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

```
ggplot(wage_data, aes(age, wage)) +
  geom_point(alpha = .04) +
  geom_smooth(method = lm, formula = y ~ poly(x = x, degree = 2))
```



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

We can observe that the predicted value for wage increases until approximately age 50, and then begins to decrease. We can infer that estimates for wage would begin to decrease at this stage possibly due to the effect of individuals beginning to retire. By fitting a polynomial, we assert that wage will not increase or decrease monotonically over the course of an individual's life – in this case, wage increases until a certain age is reached, where it then begins to decrease.

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)?

Statistically speaking, by moving from linear regression to polynomial regression we are extending the additivity assumption – we include higher order terms of X in the regression to expand on this assumption. From a substantive perspective, we are assuming that the relationship between our predictors and our outcome of interest is no longer linear – effectively, that for different values of X, we can expect different effects on the value of Y. To be clear, there can be transformations of predictors in a linear model to model a curved output, but in a non-linear model, for example, both the terms for X and X^2 would be included in the model.