Problem Set 1

Eric Karsten

January 17, 2020

Statistical and Machine Learning (25 points)

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?

Fundamentally, supervised learning is about making predictions using training data, and unsupervised learning is more exploratory because we don't observe the true predictive values we would need to calibrate a model. Some examples of problems we might tackle with supervised learning would be predicting classes (for example is something a car or a bicycle in a picture) or doing more regression style prediction (how do different demographics in an area predict what the income distribution there will be). On the unsupervised side of things, we might be doing something like feature selection (for example using lasso to pick out features that are useful predictors) or we might use it for some kind of clustering or binning (for example we want to create bins of oil wells that have similar production characteristics). I will now structure my discussion around the 5 questions discussed by the prompt.

In structured learning, we get to observe both the X's and the Y's and we use these observations to train and test a model of how different observations of X's will lead to different Y's. On the other hand, with unsupervised learning, we only observe the X's and we are leaving it up to our computer to make up some Y's. That is maybe we observe a bunch of features and we are leaving it up to the computer to assign them to being useful predictors or useless predictors. Maybe we are giving the computer some characteristics and asking it to assign thing to 10 different similar groups.

In structured learning, the target we are interested in is getting accurate predictions (generally) in our testing data. In unstructured learning, we can't directly test, but we are looking for our process to achieve some sort of optimality condition (maybe that the groups are as internally similar as possible and as different from one another as possible).

In structured learning, we imagine that there is some data generating process that leads to X's and Y's following some joint distribution (which we don't know and can't observe). What we are trying to uncover is a way of predicting the Y's based on observations of the X's based on the premise that they were generated in this joint way and that the X's have useful things to tell us about the Y's. In unstructured learning, we are generally doing something similar, we might imagine that there are some underlying groups in the data that even though we can't uncover them, did contribute to the X's that we observe. That is maybe there are differences in income in our population and an unsupervised clustering algorithm might be able to uncover something like different levels of education without observing them because these groups have dramatically different levels of income.

In structured learning, our goal in approaching the data is to understand how to get good predictions out of our X's in order to be able to predict our Y values in the future when we only observe the X's. In unstructured learning, we are trying to explore the data and uncover structures in it that might be useful for our analysis.

In structured learning, we are imagining that our model is approximating the true conditional distribution in a useful way for prediction. In unstructured learning, we are imagining that learning is the machine getting better and better at understanding underlying structures in the data.

Linear Regression Regression (35 points)

Problem Statement

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?

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

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

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?

Solution

```
# Loading Packages
library(tidyverse)
library(stargazer)

# Pretty table courtesy of
# Hlavac, Marek (2018).
# stargazer: Well-Formatted Regression and Summary Statistics Tables.
# R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

# Loading Data
data("mtcars")

# Running Regressions
m1 <- lm(mpg ~ cyl, data = mtcars)
m2 <- lm(mpg ~ cyl + wt, data = mtcars)
m3 <- lm(mpg ~ cyl * wt, data = mtcars)

# Plotting Regressions
stargazer(m1,m2,m3, omit.stat = c("f", "ser"),
          header = F,
          covariate.labels = c("Cylinder Count", "Weight (Tons)",
                               "Cylinder:Weight Interaction", "Constant"),
          dep.var.caption = "MPG",
          dep.var.labels.include = F,
          table.placement = "H",
          column.sep.width = "0pt",
          title = "Predicting MPG from Cylinder count and Vehicle Weight")
```

Table 1: Predicting MPG from Cylinder count and Vehicle Weight

	MPG		
	(1)	(2)	(3)
Cylinder Count	−2.876***	−1.508***	−3.803***
	(0.322)	(0.415)	(1.005)
Weight (Tons)		−3.191***	−8.656***
		(0.757)	(2.320)
Cylinder:Weight Interaction			0.808**
			(0.327)
Constant	37.885***	39.686***	54.307***
	(2.074)	(1.715)	(6.128)
Observations	32	32	32
R^2	0.726	0.830	0.861
Adjusted R^2	0.717	0.819	0.846
Note:			*p<0.1; **p<0.05; ***p<0.01

The regression output for parts (a), (c), and (d) is included in the table above.

a. To provide an interpretation of the regression in (a), a car with no cylinders (an ill-defined notion) would have about 38 miles per gallon, and then each added cylinder reduces that efficiency of the car by about 2.8 miles per gallon.

b. The model that is being fit in the first regression is

$$mpg_i = \beta_0 + \beta_1 cyl_i + \varepsilon_i$$

where ε_i is a mean zero error term that is independent of the number of cylinders.

c. Regression (2) in the table above reflects a specification where we add in a coefficient for weight. We notice that the cylinder count coefficient decreases when we add the weight coefficient, an indication that weight and cylinder count move together to some extent. We also note that both coefficients are negative. This means that cars with more cylinders tend to be less efficient and that the marginal cylinder reduces MPG by about 1.5. We also see that cars that weight more tend to be less efficient, that is a car that is one ton heavier than a car with an equal number of cylinders will tend to be 3 MPG less efficient. The intercept coefficient increases slightly under this specification.

d. Regression (3) in the table above reflects the specification with an interaction term. By including the interaction term, we are asserting that there is something non-linear in the effects of cylinder count and weight on vehicle mileage. That is, we are making the claim that the effect of increased weight and increased cylinder count may not move independently of one another, but rather that the magnitude of the effect of an increase in cylinder count on mileage depends on the weight of the vehicle and vice versa. We notice when we include this term in the regression that the cinder count effect and weight effect increase dramatically in magnitude (but remain negative). The interaction term is positive indicating that heavier cars will have a smaller reduction in MPG due to an increase in cylinder count and than lighter cars would (or that cars with more cylinders will have less of a reduction in MPG due to an increase in weight than those with fewer cylinders would have). We also notice that in reaction to the increase in the magnitude of the slope coefficients relative to prior regressions, this regression has a much large intercept coefficient. This is reasonable because the intercept is the centering term, but

it is a centering term for a car weighing nothing and with no cylinders, so it does not really have a well-defined interpretation since such a car doesn't exist.

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.

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

w1 <- lm(wage ~ age, data = wage)
w2 <- lm(wage ~ age + I(age^2), data = wage)

# Plotting Regressions
stargazer(w1, w2, omit.stat = c("f", "ser"),
          header = F,
          covariate.labels = c("Age", "Age$^2$", "Constant"),
          dep.var.caption = "Wage",
          dep.var.labels.include = F,
          table.placement = "H",
          column.sep.width = "0pt",
          title = "Predicting Wages as a function of Age")
```

Table 2: Predicting Wages as a function of Age

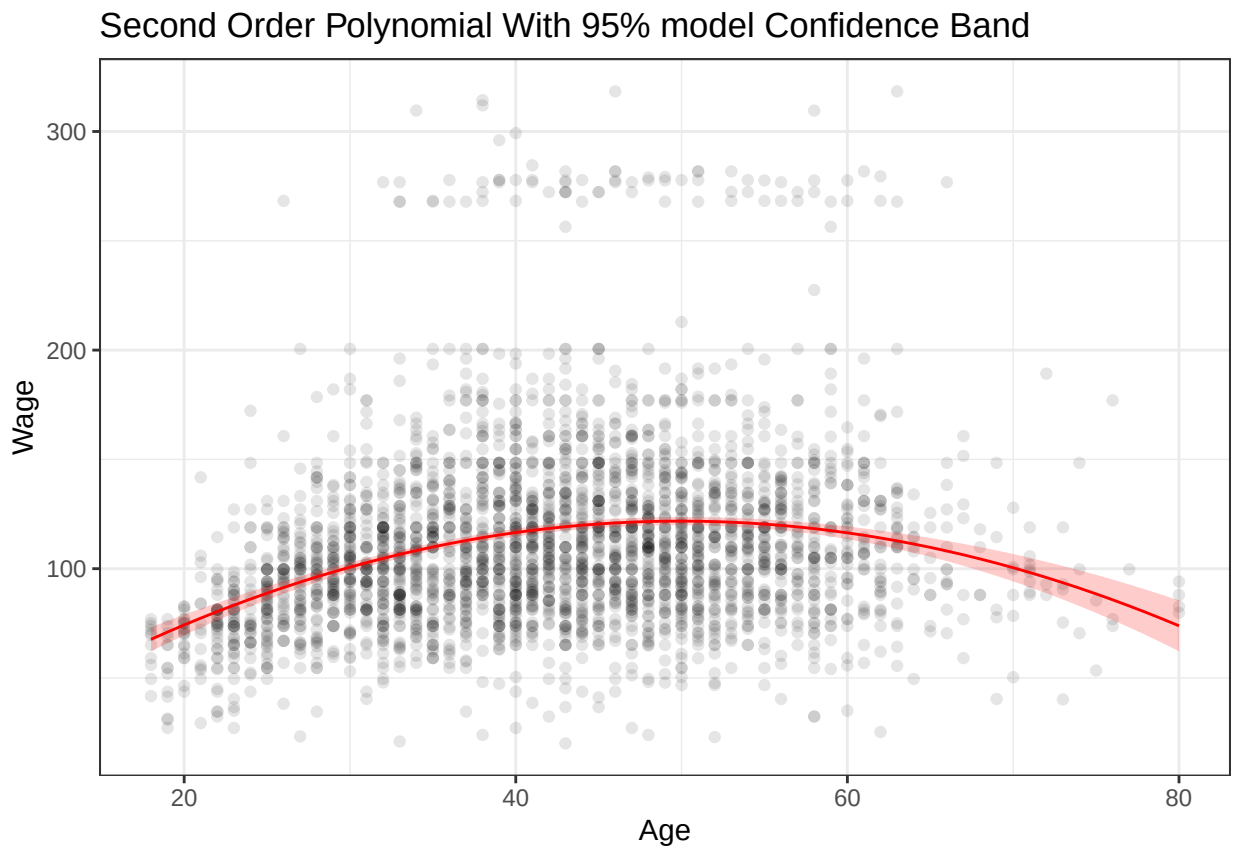
	Wage	
	(1)	(2)
Age	0.707***	5.294***
	(0.065)	(0.389)
Age^2		−0.053***
		(0.004)
Constant	81.705***	−10.425
	(2.846)	(8.190)
Observations	3,000	3,000
R^2	0.038	0.082
Adjusted R^2	0.038	0.081
Note:	*p<0.1; **p<0.05; ***p<0.01	

In the table above, equation (2) fits a second order polynomial for wages as a function of age. We see that the coefficients for both Age and Age^2 are significantly different from zero. Additionally, we see that there is a serious bump in predictive power from a naive linear model (equation (1) reported as a reference). That said, the function overall doesn't have a lot of predictive power, explaining only 8% of the variance in wages. This isn't surprising because we know that structurally, there is a lot more that goes into wage than just age, so we wouldn't expect to get fantastic predictive power from such a simple and plainly incorrect model.

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

```
# Using Predict Function to give us confidence intervals:
predictions <-
  predict(w2, newdata = tibble(age = 18:80), interval = 'confidence') %>%
  as_tibble() %>%
  mutate(age = 18:80)

# Plotting the function with confidence intervals as well as original data
wage %>%
  ggplot(aes(x = age, y = wage)) +
  geom_point(alpha = .1) +
  geom_line(data = predictions, aes(x = age, y = fit), color = "red") +
  geom_ribbon(data = predictions, aes(x = age, y = fit, ymin = lwr, ymax = upr),
              alpha = .2, fill = "red") +
  theme_bw() +
  labs(x = "Age",
       y = "Wage",
       title = "Second Order Polynomial With 95% model Confidence Band")
```



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

In the figure above, we see that our model isn't doing an particularly amazing job of fitting the data. As I have already discussed, there are a lot of things that affect wages another than age, and so it's natural that we see some unexplained within-age variation. We see above that the model slopes upwards early in peoples careers, peaking around the age of 50 and then declining through retirement. This model is somewhat imperfect in that it assumes (by being quadratic and thus symmetrical) that wages decline in retirement in the same way they rise during the early career stage. This is not necessarily a plausible assumption of the functional form

of the polynomial regression. We might obtain a better prediction by using a regression discontinuity design at the retirement age of 65, thus bringing new information into our model that is informed by the structure of the world. Additionally, we see that the error band on our estimate widens out in the early career and in the late career. This makes sense because there are fewer data points in this area to tightly calibrate the model. A finial note is that we are plotting a 95% confidence band for the model, not a 95% confidence band for the data (which would be much much wider given the lack of predictive power of age).

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)?

The most obvious difference between to two models is one of flexibility to fit the data. By virtue of having a non-linear parameter, the polynomial regression is able to fit a curve closer to data that is clearly not purely linear. A linear model is of course constrained to being a line. Statistically, the linear model is easier to interpret because we can say the coefficient represents the average earnings bump every year. On the other hand, the coefficients in the non-linear regression don't have as clean estimates. On the other hand non-linear regression can have higher predictive power (assuming it has not been over-fit to the data), so the kind of model you fit depends on what you are trying to do. If you want to interpret average structural parameters, the linear regression is a good tool, on the other hand, if you want good predictions of the effect of age and the parameter of interest is something else, then a model that is non-linear in age is superior because it won't spit out absurd estimates for older workers.