ML PS

borui sun

1/17/2020

Statistical and Machine Learning (25 points)

1. Describe in 500-800 words the difference between supervised and unsupervised learning. Supervised and unsupervised are the two main tasks within the field of machine learning. One of the major distinctions between the two types of machine learning is coping with different datasets. Supervised machine learning works with well-labeled data where both input and output variables are provided. In other words, each set of predictor feature measurements X_i is paired with an associated response measurement Y_i. However, in unsupervised machine learning, only the predictors are available, and the associated response measurement Y_i is unknown.

The difference in the forms of datasets used fundamentally determines that the goals and methods of the two types of machine learning also differ from each other. In supervised machine learning, the labeled data is used to train machines to derive the optimal hypothesis function which maps the input variables into output variables in order to better understand the "true" population function. The mapping function can also be used to predict output variables when new input data are provided. On the other hand, in unsupervised machine learning, as there is no designated outcome or target to predict, the machines are "teaching themselves" to discover the underlying structure, distribution or pattern of the dataset if there exists one. The algorithms are used against unlabeled data directly without supervision. Therefore, from a computational perspective, unsupervised machine learning can be more complex than supervised machine learning.

Moreover, the applications of the two types of machine learning in solving real world problems are different as well. Supervised machines learning is frequently used in the field of classification and regression problems depending on whether the dataset is qualitative or quantitative. Two most common examples are spam mail detection and housing price prediction. However, unsupervised machine learning methods cannot be directly applied to a regression or classification problem due to the lack of output values. Unsupervised machines learning is often seen in clustering and association problems. Clustering problems can be interpreted as finding innate clusters that may exist in the data- for example, categorization of customers into different groups based on their purchase histories, while association often refers to the problem of unveiling the connections among data objects inside large databases. Another example would be promoted web advertising based on one's browsing histories.

It is also important to address the deficiencies associated with each of the two types of machine learning. One of the biggest challenges of unsupervised machine learning is the difficulty to measure the accuracy of its algorithm, as there is training dataset that one can easily compare its prediction with target value and readjust its model according. Sometimes additional work may be required in order to acquire meaningful results. Therefore, unsupervised machine learning can be less predictable than supervised machine learning. While accuracy measurements are extremely helpful in model evaluation, labeled data often require enomorous amount of human resources and are still difficult to find in many research or industrial fields. Accesses to unlabeled data are much more attainable, which leads to huge demand for unsupervised machine learning.

Linear 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?

The output in our model is miles per gallon (mpg). We have two model parameters: the constant term is 37.88 and the cofficient for cyl is -2.88 as shown below. The 37.88 (β_0 or intercept) is the expected mean value of miles per gallon when a given car has 0 cylinders. β_1 (coefficient on cyl) indicate that 1 unit increase in the number of cylinders in a given car is associated with 2.88 miles/(US) gallon decrease in the car's speed.

```
ols_biva <- lm(mpg~cyl, data = mtcars);tidy(ols_biva) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	37.88458	2.0738436	18.267808	0
cyl	-2.87579	0.3224089	-8.919699	0

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

$mpg = \beta_0 + \beta_1 cyl$ where $\beta_0 = 37.88$ and $\beta_1 = -2.88$ in this dataset.

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

After adding vehicle weight (wt) to the specification, the magnitude of the coefficient on cyl decreases by 1.37 but the coefficient still remains negative and statistically significant. More importantly, -1.5 now becomes the estimate effect of cyl on mpg when holding wt constant. The intercept also increases from 37.88 to 39.69. The coefficient on our new regressor, vehicle weight, is negative and statistically significant. We can also find that the adjusted-R^2 also increased after adding vechile weight into our model, indicating that more variance for our dependent variable (mpg) is explained by the independent variables after controling for wt.

```
ols_multi <- lm(mpg~cyl + wt, data = mtcars);tidy(ols_multi) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	39.686262	1.7149840	23.140893	0.0000000
cyl	-1.507795	0.4146883	-3.635972	0.0010643
wt	-3.190972	0.7569065	-4.215808	0.0002220

```
ols_biva %>% glance() %>%
  bind_rows(glance(ols_multi)) %>%
  mutate(model = c('bivariate OLS', 'multivariate OLS')) %>%
  select(12, 1:6)%>%kable()
```

model	r.squared	adj.r.squared	sigma	statistic	p.value	df
bivariate OLS	0.7261800	0.7170527	3.205902	79.56103	0	2
multivariate OLS	0.8302274	0.8185189	2.567516	70.90836	0	3

- 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?

Aftering including an interaction term between weight and cylinders into our model, the magnitude of

coefficients on cyl and wt increases while their estimates are still negative and statistically significant.The interaction term is also statistically significant at 0.05 level. The constant term also increases. Including the interaction term increases the adjusted-R^2 by a small amount (~0.01).

More importantly, the interpretions of our coefficients on cyl and wy also changed. By including the interaction term in our function, we are theorectically asserting that the effect of cyl (or wt) on mpg is conditional on the value of wt (or cyl). In other words, the estimated effect of cyl and wt on mpg is no longer constant.

```
ols_inter <- lm(mpg ~ cyl*wt, data = mtcars); tidy(ols_inter) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	54.3068062	6.127535	8.862749	0.0000000
cyl	-3.8032187	1.005028	-3.784193	0.0007472
wt	-8.6555590	2.320122	-3.730648	0.0008610
cyl:wt	0.8083947	0.327322	2.469723	0.0198824

```
glance(ols_inter)[, 1:6] %>% kable()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.8605954	0.8456592	2.367761	57.61805	0	4

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.).

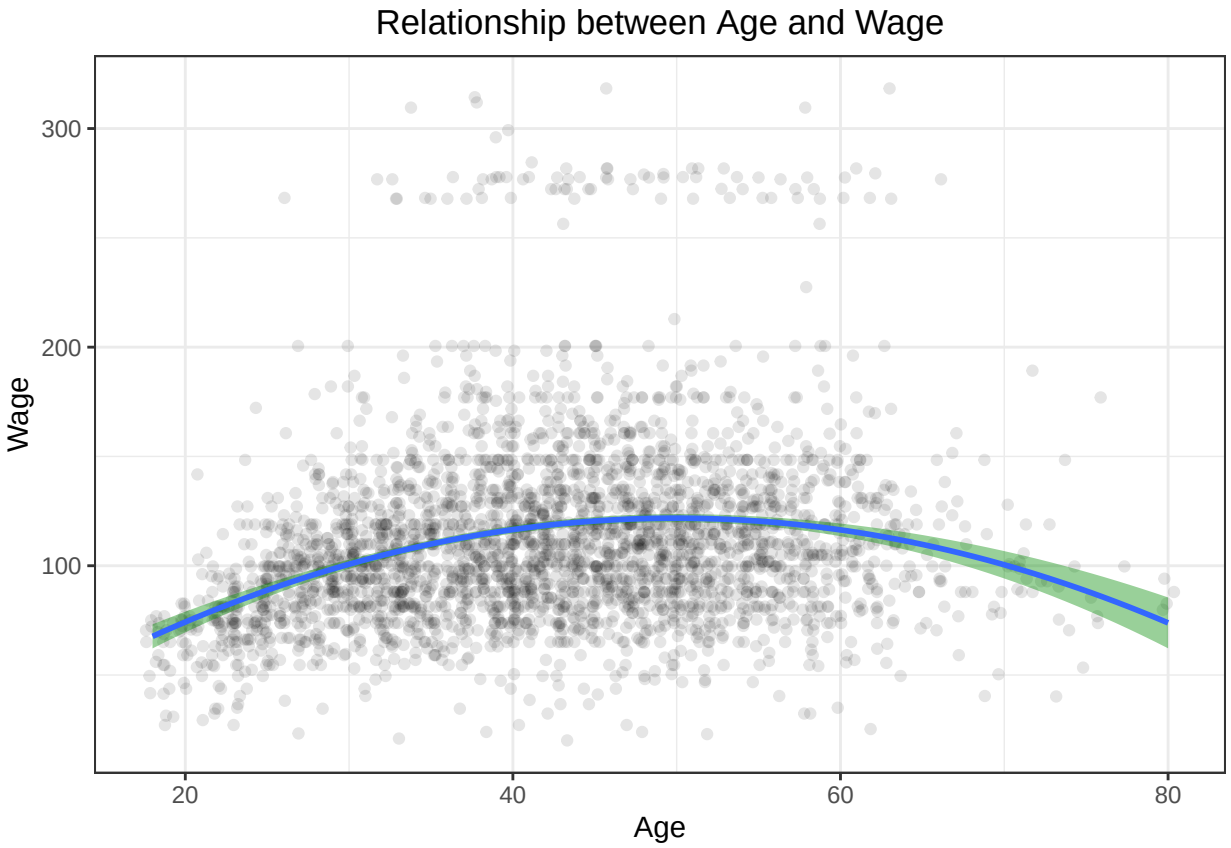
As shown below, the regression coefficients for age and age^2 are 5.29 and -0.05. As both estimates are significant at 0.01 level, it implies that the effect of age on wage level is constantly changing as age increases. The negative sign of the quadratic term indicates a concave relationship between age and wage. That is, there is a maximum amount of wage that one can earn during his/her lifetime.

```
polynomial <- lm(wage ~ age + I(age^2), data = wage_data);tidy(polynomial) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-10.4252243	8.1897803	-1.272955	0.2031326
age	5.2940300	0.3886886	13.620236	0.0000000
I(age^2)	-0.0530051	0.0044318	-11.960103	0.0000000
+ b. (10) Pl	ot the functio	n with 95% c	onfidence int	erval bounds.

```
ggplot(wage_data, aes(x = age, y = wage)) +
  geom_jitter(alpha = 0.1) +
  geom_smooth(fill = "#008B00", method = "lm", formula = y~poly(x,2), se = TRUE, level = 0.95) +
  labs(title = "Relationship between Age and Wage",
```

```
       x = "Age",
       y = "Wage")
```



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

In the above figure, we can see that the regression line is concave instead of a linear. As a person's age increases, his/her wage also increase, but as the person's age reaches a certain level (50-ish in this case), his/her wage reaches a maximum point and starts to decrease. Fitting a polynomial regression implies that we assume the relationship between age and wage is polynomial.

- 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)?

While a polynomial regression is linear in the parameters, the underlying relationship is polynomial (non-linear). Linear regression, on the contrary, assumes that the underlying relationship is linear. Also, the effect of independent variable on outcome is constant in a linear regression but is constantly changing in a polynomial regression. Also, while polynomial (non-linear) regression is more flexible than linear and can fit into a variety of curves, its estimates are more complex to interpret.