MACS33002: Homework #1

Yujiao Song

Due Friday, Jan 17 by 5pm

##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 learning: ##For each there is a set of variables that might be denoted as inputs, which are measured or preset. These have some influence on one or more outputs. For each example the goal is to use the inputs to predict the values of the outputs. This exercise is called supervised learning. Supervised learning is the learning of the model where with input variable (say, x) and an output variable (say, Y) and an algorithm to map the input to the output. It is called supervised learning because the process of an learning(from the training dataset) can be thought of as a teacher who is supervising the entire learning process. Thus, the "learning algorithm" iteratively makes predictions on the training data and is corrected by the "teacher", and the learning stops when the algorithm achieves an acceptable level of performance(or the desired accuracy).That is, Y = f(X) There are two simple but powerful prediction methods in supervised learning: the linear model fit by least squares and the k-nearest-neighbor prediction rule. With supervised learning there is a clear measure of success, or lack thereof, that can be used to judge adequacy in particular situations and to compare the effectiveness of different methods over various situations. There are four major issues to consider in supervised learning: bias-variance tradeoff, function complexity and amount of training data, dimensionality of the input space, noise in the output values. ##Unsupervised learning: ##Unsupervised learning is where only the input data (say, X) is present and no corresponding output variable is there. It is also known as self-organization and allows modeling probability densities of given inputs. Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group. A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses many other domains involving summarizing and explaining data features. Generative adversarial networks can also be used with unsupervised learning, though they can also be applied to supervised and reinforcement techniques.

##The difference between supervised and unsupervised learning in process is in a supervised learn-

ing model, input and output variables will be given while In unsupervised learning model, only input data will be given. The algorithms for input data in supervised learning are trained using labeled data while in unsupervised learning algorithms are used against data which is not labeled. In data generating process, supervised learning model uses training data to learn a link between the input and the outputs, unsupervised learning does not use output data. Supervised learning allows you to collect data or produce a data output from the previous experience. Unsupervised machine learning helps you to finds all kind of unknown patterns in data. Regression and Classification are two types of supervised machine learning techniques. Clustering and Association are two types of Unsupervised learning

##linear regression

```
data(mtcars)
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"
## [10] "gear" "carb"
```

```
lmmpg=lm(mpg~cyl,data=mtcars)
summary(lmmpg)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.981 -2.119  0.222  1.072  7.519
##
## Coefficients:
##             Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   37.885      2.074   18.27       < 2e-16 ***
## cyl           -2.876      0.322   -8.92 0.00000000061 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.21 on 30 degrees of freedom
## Multiple R-squared:  0.726,  Adjusted R-squared:  0.717
## F-statistic: 79.6 on 1 and 30 DF,  p-value: 0.000000000611
```

##1(a) Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model? ##The intercept value is 37.8846 and the parameter value is-2.8758.The linear equation is mpg = 37.88 - 2.88*cyl. The parameter equals -2.88 and its p value is smaller than 0.05. Thus, cyl is a significant variable. ##1(b) Write the statistical form of the simple model in the previous question (i.e., what is the population regression function?). ##population form: mpg=-2.8758cyl+37.8846

```r
lmmpg=lm(mpg~cyl+wt,data=mtcars)
summary(lmmpg)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.289 -1.551 -0.468  1.574  6.100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.686      1.715   23.14  < 2e-16 ***
## cyl           -1.508      0.415   -3.64  0.00106 **
## wt            -3.191      0.757   -4.22  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.57 on 29 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.819
## F-statistic: 70.9 on 2 and 29 DF,  p-value: 0.00000000000681
```

##1(c) Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc. ##The intercept value change to 39.6863 and the parameter on cyl change to -1.5078 which the absolute value is smaller when compared with the condition of mpg is only depend on cyl.The effect of cyl on mpg is less. The parameter on wt is -3.1910. All of the p-values of parameters are small than 0.05. Thus, both cyl and wt are significant variables.

```r
lmmpg=lm(mpg~cyl+wt+cyl*wt,data=mtcars)
summary(lmmpg)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + cyl * wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.229 -1.350 -0.504  1.465  5.234
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)
## (Intercept)   54.307      6.128    8.86 0.0000000013 ***
## cyl           -3.803      1.005   -3.78      0.00075 ***
## wt            -8.656      2.320   -3.73      0.00086 ***
```

```
## cyl:wt         0.808      0.327    2.47      0.01988 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.37 on 28 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.846
## F-statistic: 57.6 on 3 and 28 DF,  p-value: 0.00000000000423
```

##1(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? ##All of the coefficients are significant and they are different from the previous two models. This means the mutiplicatiove interaction term has an effect on mpg.

##nonlinear regression

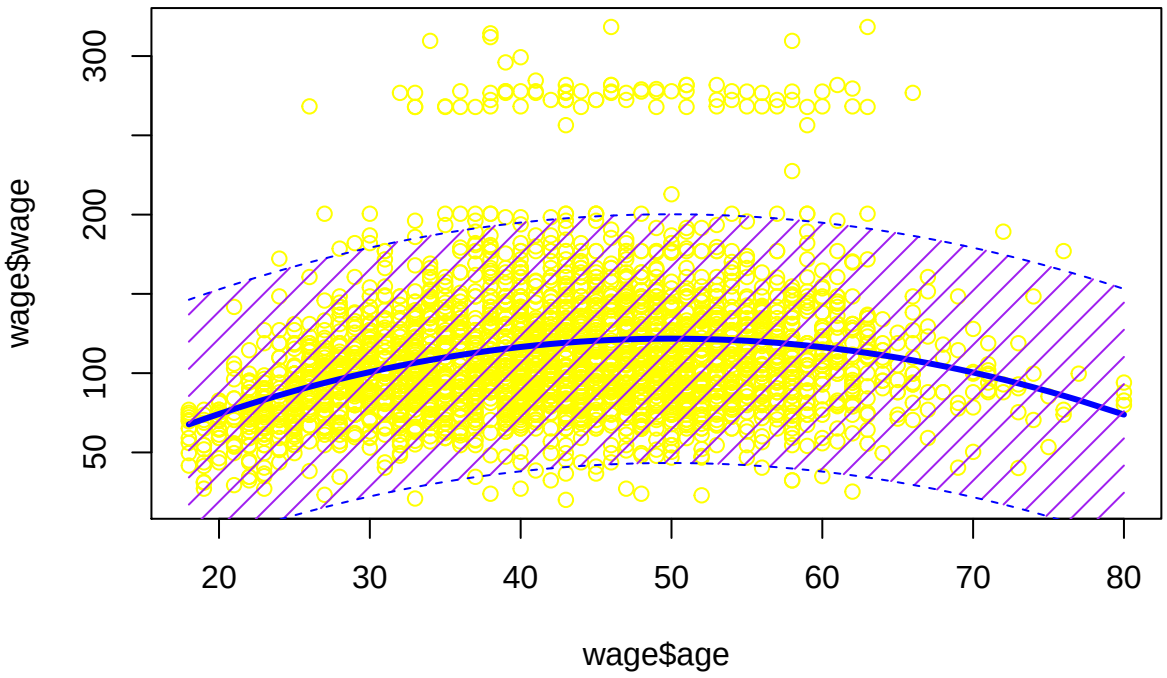
```
wage <- read.csv('wage_data.csv', header=T,sep=',')
lmwage=lm(wage~age+I(age^2),data=wage)
summary(lmwage)
```

```
##
## Call:
## lm(formula = wage ~ age + I(age^2), data = wage)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -99.13 -24.31  -5.02  15.49 205.62
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.42522    8.18978   -1.27      0.2
## age           5.29403    0.38869   13.62   <2e-16 ***
## I(age^2)     -0.05301    0.00443  -11.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40 on 2997 degrees of freedom
## Multiple R-squared:  0.0821, Adjusted R-squared:  0.0815
## F-statistic:  134 on 2 and 2997 DF,  p-value: <2e-16
```

##1(a) 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.). ##The parameter on age is 5.294030, the parameter on age^2 is -0.053005 and the intercept is -10.425224. All of the coefficient are significant since they have small p-value.

```
plot(wage$age, wage$wage, col="yellow")
new.age = seq(min(wage$age), max(wage$age), length.out=100)
preds <- predict(lmwage, newdata = data.frame(age=new.age), interval = 'prediction')
```

```
lines(sort(wage$age), fitted(lmwage)[order(wage$age)], col="blue", type = "l", lwd = 3)
polygon(c(rev(new.age), new.age), c(rev(preds[ ,3]), preds[ ,2]), density=10, col = 'purple
lines(new.age, preds[ ,3], lty = 'dashed', col = 'blue')
lines(new.age, preds[ ,2], lty = 'dashed', col = 'blue')
```



##b. (10) Plot the function with 95% confidence interval bounds. ##c. (10) Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression? ## We can see from the plot that there are lots of outliers. However, most of our points are included in the 95% confidence interval means that this polynomial did a good job in explaining the variaility. ##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 is easier to use, simpler to interpret, and you obtain more statistics that help you assess the model. While linear regression can model curves, it is relatively restricted in the shapes of the curves that it can fit. Sometimes it can't fit the specific curve in your data.

##Nonlinear regression can fit many more types of curves, but it can require more effort both to find the best fit and to interpret the role of the independent variables. Additionally, R-squared is not valid for nonlinear regression, and it is impossible to calculate p-values for the parameter estimates.