# MACS 33002: PSET 1

*Deniz Turkcapar*

## Statistics and Machine Learning

### Question 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. . .

In supervised learning, we have data that is fully labeled that we use to train the algorithm. The labels show that each example in the training dataset is tagged with the correct answer we hope the algorithm should come up on its own. The relationship between X and Y is as follows: we have input variables (X) and output variable (Y). We use an algorithm to learn the specific mapping function from the input to the output. This can be summarized simply like this: Y = f(x). The objective of supervised learning is to approximate the mapping function as good as possible so that when we have new input data points (X), we can predict the most accurate output variables (Y) for the given data. Hence, the target that we are interested in is having the most accurate mapping function that predicts output (Y) as correctly as possible. Learning is conceptualized by the mapping relationship that X and Y has in the training dataset. There are 2 main areas where supervised learning is useful: classification and regression. The goal of classification is to ask the algorithm to predict a discrete value and identify the input data as the member of a particular class by using the relationship observed in the training dataset. Regression problems look at continuous data to predict a possible outcome Y. It asks the following question: "given a particular value of x, what is the expected value of the Y variable?". Supervised learning is suitable for problems in which we have a set of available reference points that serve as the ground truth to train the algorithm. In the real world, however, it might be hard to find perfectly labeled data.

Unsupervised learning makes use of a deep learning model that is given a dataset without explicit instructions on what to do with it. In this case, the training dataset contains collections of examples without a specific desired outcome. In unsupervised learning, the relationship between X and Y is quite different: we do not have a Y. We have a lot of input datapoints (X) only and no corresponding output variables (Y). The goal of unsupervised learning is to model the unknown structure or distribution of data to learn more about it. Unlike supervised learning, there is no set of correct answers and no prescribed algorithm as a teacher. Therefore, we only have a set of features and no targets. The algorithm is left on its own to figure out the interesting structures in the data. Unsupervised learning algorithms can be further categorized into addressing two components that help to conceptualize the learning process: clustering and association. A clustering problem is where we want to discover the underlying groupings in the data. Association is where we want to discover rules that describe a significant portion of the data, which can also explain the trends we see in the data. One problem that can happen in the data generating process is that if there is a lack of examples, the model will fail to generalize well and this causes overfitting because the empirical distribution is not entirely representative of the data generating process.

## Linear Regression

### Question 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?**

```r
wageData <- read.csv("wage_data.csv")
```

```r
model <- lm(mpg ~ cyl, data=mtcars)
summary(model)
```

```

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
```

The coefficient associated with cyl is estimated to be -2.88, with a standard error of 0.32. The result of the two-tailed t-test is significant, implying that the coefficient is significantly nonzero.

The intercept is found to be 37.88 with a standard error of 2.07, which is also statistically significant. We get an R-squared value of 0.73 which is relatively high explained variance. The residuals seem to be distributed evenly.

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

Y $= \beta_0 + \beta_1 X + \epsilon$ where Y is mpg and X is cyl

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

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

```
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
```

The coefficient associated with cyl is estimated to be -1.51, with a standard error of 0.41. The coefficient is lower in magnitude than the regression before, implying that there might be a correlation between cyl and wt. The result of the two-tailed t-test is still significant, implying that the coefficient is significantly nonzero.

The coefficient associated with wt is estimated to be -3.19, with a standard error of 0.76. The result of the two-tailed t-test is significant, implying that the coefficient is significantly nonzero.

The intercept is found to be 39.69 with a standard error of 1.72, which is also statistically significant.

We get an R-squared value of 0.83 which is relatively high explained variance. This is higher than the 0.73 we got earlier, meaning that the two variable-model has higher explanatory power. The residuals seem to be distributed evenly.

### (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?

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

```
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
```

The interaction term is positive, meaning that there is a positive correlation in this sample between cyl and wt (which makes sense because cars with more cylinders will tend to be heavier). Due to the inclusion of the interaction term, the coefficients are more negative for both variables, compared to the 2 previous regressions. The interaction wterm would assert that we suspect that the magnitude of the effect of wt on mpg will change as cyl changes (and that the magnitude of the effect of cyl on mpg will change as wt changes).

# Nonlinear Regression

## Question 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.).

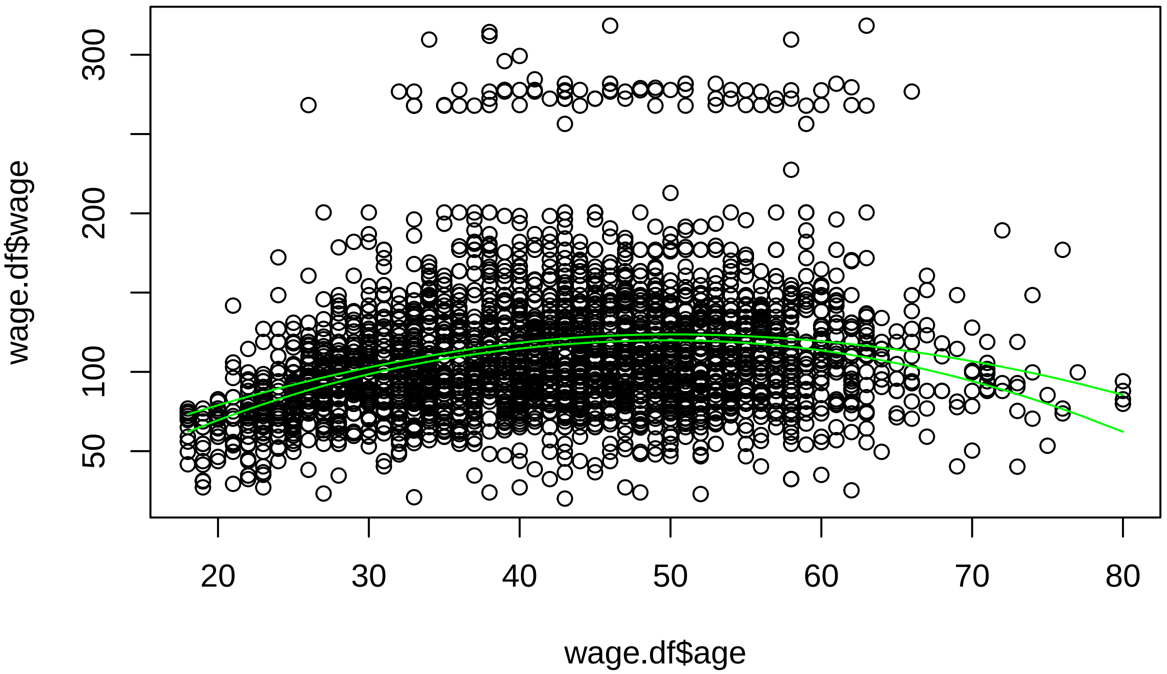
```
wage.df <- read.csv('wage_data.csv')
wage.df <- wage.df[order(wage.df$age),]
model <- lm(wage~poly(age,2,raw=T),data=wage.df)
summary(model)
```

```

Call:
lm(formula = wage ~ poly(age, 2, raw = T), data = wage.df)

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    
poly(age, 2, raw = T)1   5.29403    0.38869   13.62   <2e-16 ***
poly(age, 2, raw = T)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
```

The linear term has coefficient 5.29 and the quadratic term has coefficient -0.053, both of which are statistically significant. These coefficients tell us that as age initially increases the wage also increases, but as it increases even more, the impact of the quadratic term dominates and the estimated wage starts decreasing.

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

```
p <- predict(model,wage.df,interval='confidence',level=0.95)
plot(wage.df$age,wage.df$wage)
lines(wage.df$age,p[,3],col='green')
lines(wage.df$age,p[,2],col='green')
```



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

We can observe that a unit increase in age causes a 5% increase in wage. We can observe a negative correlation between the age squared and the wage. On the other hand, there is a positive correlation between age and wage. In the 95% confident interval fitted plot, we can see that there is a lot of variability in the data with some clear outliers at the top of the plot, signifying that those outliers contain very high numbers. By fitting a polynomial model, we assert that there is a relationship that is more complex than a linear regression. We've done polynomial regression because we thought that the linear regression model was not adequte.

**(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 model is useful for modeling the relationship between a scalar dependent variable and one or more independent variables in a linear relationship. Polynomial regression is a form of linear regression where higher order powers (2nd, 3rd power etc.) of an independent variable is included in explaining the relationship between the independent and dependent variables. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y (denoted $E(y|x)$). Even though a polynomial regression fits a nonlinear model to the data, it is linear as a statistical estimation problem as the regression function $E(y|x)$ is linear in the unknown parameters that are estimated from the data. Advantages of polynomial regression include: polynomial providees the best approximarion of the relationship between the X and Y variables, polynomial can fit a wide range of curvature, and a variety of functions can be fit under it. Disadvantages of using polynomial regression include: the presence of outliers in the data can significantly affect the results of nonlinear analysis, polynomial regression is very sensitive to outliers, and there are fewer model valudation tools for detecting outliers in nonlinear regression than there exists for linear regression.