problem-set-1

See syllabus for submission details.

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 machine learning uses a set of features X from a population n to predict some categorical or continuous outcome Y. The Y is a known quantity or quality, and therefore the accuracy of the model can be ascertained by comparing the predicted and actual Y. The target is to predict this quantity Y in a different testing data set. The goal in approaching this data is to learn how these X features interact with each other and the outcome Y.

Unsupervised machine learning uses a set of features X from a population n to predict some latent characteristic Y. This latent characteristic is usually group membership into 1 of k groups. The relationship between X and Y is more complex as it is not explicitly evident in the data. The assumption is that the population n is made up of more homogeneous subgroups K, and the X features can be used to discover which of the K groups an individual belongs to. The goals is to find underlying substructure to the data that is assumed to be a pooling of several clusters or subgroups.

Linear Regression 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?

data(mtcars) head(mtcars) linearModel <- lm(mpg ~ cyl, data=mtcars)
print(linearModel) summary(linearModel)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.8846 2.0738 18.27 < 2e-16 ***cyl -2.8758 0.3224 -8.92 6.11e-10***

Multiple R-squared: 0.7262, Adjusted R-squared: 0.7171 F-statistic: 79.56 on 1 and 30 DF, p-value: 6.113e-10

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

mpg ij

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? Non-linear Regression (40 points)

library(foreign) wage_data <- read.csv("C:/Users/sbhavani/Desktop/wage_data.csv")

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.).

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?

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)?