HOMEWORK 1 MATH 484-564, REGRESSION

DUE SEPTEMBER 6TH 2024, FRIDAY, 11:59PM

(1) (6 points) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n, the number of data; and p, the number of predictors/features in each of the scenario.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

(2) (6 points) What are the advantages and disadvantages of parametric and non-parametric methods?

3. (6 points) In practice, we usually don't calculate least squares lines manually; instead, we use statistical software like R to handle the computations. However, in this exercise, we will manually compute the least squares line using the formula

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2},$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

to understand the process better. (Note: There are others equivalent way of writing the above two formulas.)

Given the following data points:

x	y
1	10
2	14
4	12
6	13
7	15
8	12
10	13

Calculate the least squares line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$, without using software such as R, Excel, Python and so on, a typical hand-held calculator is fine.

Write your answer in fraction, do not simplified to decimals.

4. (6 points) In Problem 3, notice that with $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ we have the following:

x	y	$\widehat{y}$	$y - \widehat{y}$
1	10	5303/446	-843/446
2	14	2693/223	429/223
4	12	2776/223	-100/223
6	13	2859/223	40/223
7	15	5801/446	889/446
8	12	2942/223	-266/223
10	13	3025/223	-126/223

Take note that the values in the 4th column sum to zero!

This is actually not a coincidence, in fact, show that for the least squares line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ obtained from the training data $(x_1, y_1), \cdots, (x_n, y_n)$, we have

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0.$$

5. (6 points) Consider a statistical learning model trained to predict a continuous outcome Y based on features X. The model is trained on a finite training dataset, and we are interested in understanding the tradeoff between variance and bias in the model's predictions.

   (a) Explain what is meant by "bias" and "variance" in the bias-variance tradeoff in the context of statistical learning models.

   (b) Suppose you have the following scenarios:

   - **Scenario 1:** A model that is very simple, like a linear regression model with only one feature.
   - **Scenario 2:** A more complex model, like a polynomial regression model with many features.

   (i) Which scenario is more likely to have high bias and low variance? Explain.

   (ii) Which scenario is more likely to have low bias and high variance? Explain.