CME 250 Homework 1

Instructor: Sherrie Wang (sherwang@stanford.edu)

Available: Wednesday, January 16
Due: Friday, January 25, 5:00pm

Please submit Part 1 of the homework via Google Forms, and submit Parts 2 and 3 below via Gradescope.

For each question, clearly identify your final answer and show any work or code used to arrive at the answer. Answers without the corresponding code or justification will not receive credit. You may use any programming language of your choice.

Part 2. Applied Exercise

Question 1.

In this question, we are going to create synthetic data so that we know the underlying data generation process, and then try to recover the real relationship between x and y using three different linear regression models.

(a) [2 points]

Generate a synthetic dataset of one hundred $(x^{(i)}, y^{(i)})$ samples as follows.

1. Draw 100 x values uniformly at random from the interval $[0, 10]$.
2. Compute $y = 0.03x^3 - 0.3x^2 + 0.3x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ is the irreducible error term. In other words, y is a cubic function of x with noise added from a Gaussian distribution.

Plot your dataset on the domain $x \in [0, 10]$ and for the relevant range. (No need to print out the actual dataset in your solution.)

Hint: If you are using Python, the functions in the `numpy.random` routine may be useful. In R, check out `runif` and `rnorm`.

(b) [2 points]

Generate 10 such synthetic datasets and run simple linear regression on each dataset. That is, your model is

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

Plot the linear equation computed from one run on the same plot as the dataset. Report the mean and standard deviation of the R^2 across the 10 runs to 2 decimal places.

(c) [2 points]

For each of 10 synthetic datasets, run a multiple linear regression with the features x, x^2, x^3. That is, your model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{2}$$

Plot the regression curve computed from one run on the same plot as the dataset. Report the mean and standard deviation of the R^2 across the 10 runs to 2 decimal places.

(d) [2 points]

For each of 10 synthetic datasets, run a multiple linear regression with the features $x, x^2, x^3, \ldots, x^{10}$. That is, your model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{10} x^{10} + \epsilon \tag{3}$$

Plot the regression curve computed from one run on the same plot as the dataset. Report the mean and standard deviation of the R^2 across the 10 runs to 2 decimal places.

(e) [2 points]

Which model from parts (b)–(d) has the most bias? The most variance? Which model would you expect to generalize the best to unseen data?

Part 3.

Question 2. Young People Survey

Download the `responses.csv` dataset from Kaggle Datasets at:
`https://www.kaggle.com/miroslavsabo/young-people-survey`

We will be playing with this dataset for all 4 homework assignments. In particular, we are going to predict the `Highest education achieved` variable using the other features. To read more about what each feature in the dataset represents, navigate to the Overview tab at the link above.

(a) [1 point]

Read the dataset into an array or dataframe using your language of choice. Drop all rows that have any missing values (often shown as `NaN`s or `NA`s). (We will investigate strategies for handling missing data in a future homework.) What are the dimensions of this resulting matrix or dataframe?

(b) [1 point]

Create a vector called `y` from the `Education` column of the dataset, and a matrix called `X` containing all other columns.

(c) [3 points]

Ten of the features are encoded as categories, while the remaining are integers. (The response `Education` is also encoded categorically.) Transform these features into numerical or binary values as appropriate.

Hint: Look for existing classes or functions that do this in your language of choice.