COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 1

1 Directions:

- **Due: Thursday February 3, 2022 at 9pm.** Late submissions will be accepted for 24 hours after that time, with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low), then you might receive a zero or only partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2 Problems

Problem 1. [16 points; 4 each part] For each of parts (a) through (d), indicate whether we would generally expect the performance of fitted models on future data the be better or worse if the model class has many degrees of freedom (e.g. class of high degree polynomials) compared to a model class with few degrees of freedoms (e.g. constant or linear functions). Briefly explain why (1-3 sentences).

(a). The number of samples n is large, and the number of features p is small.

(b). The number of features p is large, and the number of samples n is small.

(c). The relationship between the features $\{X_1, \ldots, X_p\}$ and the response (Y) is highly non-linear.

(d). The variance of the noise terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

(a). Flexible would be better, since with lots of data, we can get accurate estimates for coefficients even for non-linear models – we can "average out the noise"

(b). Inflexible would be better, since with few observations, the flexible method might over-fit (fit the trend plus the noise), resulting in poor generalization.

(c). Flexible would be better if the relationship is highly non-linear to better capture the trend.

(d). Inflexible would be better, as the large noise could lead to over-fitting by flexible models.

Problem 2. [18 points total; 6 each part] You will now think of some real-life applications for machine learning.

(a). Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors.

(b). Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors.

(c). Describe three real-life applications in which cluster analysis might be useful.

(a). classification:

- self-driving cars, using lidar and images to classify if there is a pedestrian nearby;
  - response: binary variable of pedestrian present or not
  - predictors: pixel values from cameras
- Sports gaming– predicting which team will win the championship.
  - response: Name of a team
  - predictors: statistics collected from past games, of individual players' performances and team performances.
- Finger print reader for access to a building or system
  - response: name of the user (if authorized)
  - predictors: pressure readings from a touchpad

(b). regression

- predicting future prices (stock market or housing market),
  - response: price of a stock (or house)
  - predictors: past prices of that and related stocks and economic indicators
- predicting rainfall from a storm
  - response: rainfall in inches over a 5 day period
  - predictors: past rainfall, humidity, wind, and temperature data at location of interest and surrounding region.
- identifying appropriate medical doses based on body
  - response: # grams of medicine
  - predictors: health outcome, dosage level, health metrics (weight, height, age, blood pressure, etc)

(c). clustering

- grouping customers by purchasing habits (eg Amazon, Netflix, etc.)
- grouping genes by expression level in a population (not necessarily with different conditions) to determine later experiments for identifying if they causally interact

- grouping together workers (or population) by sleep patterns to understand if there are early bird/night owls or a spectrum

Problem 3. [10 points]

We want to learn a model to predict Y. Let n denote the number of samples of data. Using calculus, derive the optimal constant function $\widehat{Y}(X) = \beta_0^*$ under mean square error

$$\beta_0^* = \arg\min_{\beta_0 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (Y(i) - \beta_0)^2 .$$

Make sure to check that you found a minimizing β_0, not a maximizing β_0.

Take the first derivative and set it equal to 0.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\beta_0} \frac{1}{n} \sum_{i=1}^{n} (Y(i) - \beta_0)^2 \tag{1}$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2 (Y(i) - \beta_0)(-1) \tag{2}$$

$$\implies \quad 0 = \sum_{i=1}^{n} (Y(i) - \beta_0) \tag{3}$$

$$\implies \quad n\beta_0 = \sum_{i=1}^{n} Y(i) \tag{4}$$

$$\implies \quad \beta_0 = \frac{1}{n} \sum_{i=1}^{n} Y(i) \tag{5}$$

$$\tag{6}$$

So the mean of the Y values is a stationary point. Taking the second derivative,

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\beta_0} \frac{1}{n} \sum_{i=1}^{n} (Y(i) - \beta_0)^2 = \frac{\mathrm{d}}{\mathrm{d}\beta_0} \frac{1}{n} \sum_{i=1}^{n} 2 (Y(i) - \beta_0)(-1) \tag{7}$$

$$= 2 \tag{8}$$

Since the second derivative is positive for all β_0, we can conclude that $\beta_0^* = \frac{1}{n} \sum_{i=1}^{n} Y(i)$ is the global minimizer of the MSE.