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.

Better. If the sample size is large, it will be likely to have smaller MSE and if the number of feature p is small it will be less likely to over-fit.

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

Worse. Similar idea to (a), if the number of features p is large, it will be likely to overfit and small sample size leads to larger MSE because an outlier can effect more in.

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

Better. Because it has more degree of freedom, it will be more flexible and fit better.

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

Worse. It will fit to the noise and increase the variance.

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.

By using the text, it might be able to check whether the email is spam.

By using the picture of flowers, I might be able to classify the species of flower.

By using the picture of people, it might be able to recognize face which could be used to unlock cell phone for instance.

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

With relationship between parents' income and child's income, we can predict child's the income from parents' income

With relationship between temperature and the population, it might be possible to predict future population in the city effected by global warming from the temperature of the city.

With relationship between price and the number of sales, we might be able to predict the number of sales from the price and find the best price to set.

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

By using the 3D image information, it might be useful to distinguish different tissue.

By using homologous sequence, it might be useful to group the families.

By using the shopping item on the Internet, it might be useful to group items.

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 $\hat{Y}(X) = \beta^*_0$ under mean square error

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

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

=arg min 1/n $((Y(1)-\beta_0)^2 + (Y(2)-\beta_0)^2 + (Y(2)-\beta_0)^2 + \ldots + (Y(n)-\beta_0)^2)$

=arg min 1/n $((Y(1)^2+Y(2)^2+Y(3)^2+\ldots+Y(n)^2)+ 2(Y(1)+Y(2)+Y(3)+\ldots+Y(n))\,\beta_0 + n\beta_0^2)$

$=\arg\min \frac{1}{n}\left(\sum_{i=1}^{n} Y(i)^2 - 2\,\beta_0 \sum_{i=1}^{n} Y(i) + n\beta_0^{\,2}\right)$…i

$\therefore dY/d\beta_0 = \frac{1}{n}(2n\beta_0 - 2\sum_{i=1}^{n} Y(i))$

$\because$i, or quadratic function

when $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} Y(i)$ the optimal function has minimum.

$\therefore \beta_0^{\,*} = \{\frac{1}{n}\sum_{i=1}^{n} Y(i)\}$