COMS 474 HW 1

Haadi Majeed

January 31, 2022

Contents

1 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.

- When working with large amounts of data, and fitting it with a small number of features, the model that comes out of it will not accurately represent the dataset as it may be treating trends and their variations as just noise that it ignores. Having p be too small would result, generally, in under-fitting the data. This would result in a poorer model in general.

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

- When working with small amounts of data and many features, or higher flexibility, it would over-fit the data where it would effectively treat the data as if there was no noise. Such can be seen in the image below, where the model to the 16th power successfully hits every point of data, however is not very useable for future data, which is what we care about. This would result in a poorer model in general.

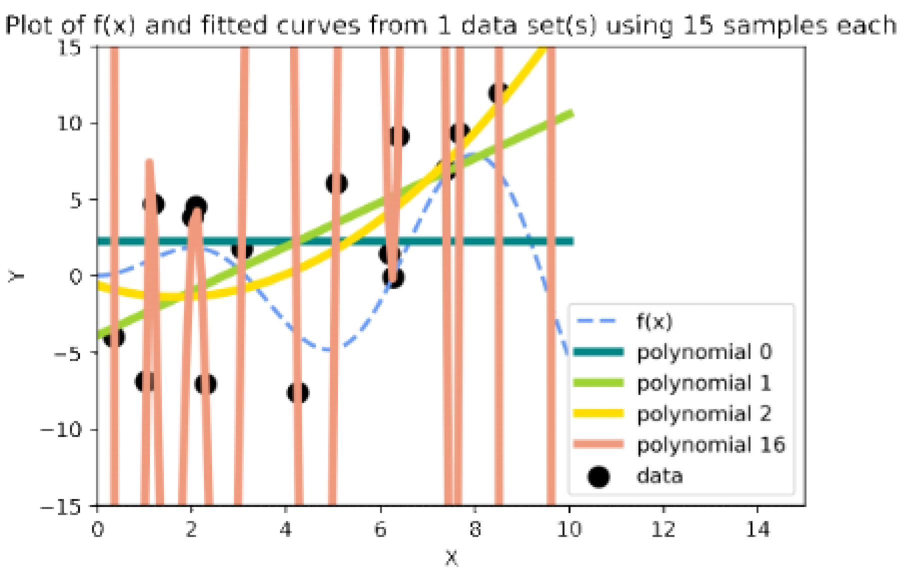

Image from page 8 of January 27th's Notes

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

- waffle

(d) The variance of the noise terms, $\{X_1...X_p\}$, is extremely high

- When the noise of the data is extremely high, it becomes more difficult to fit a model to it as the future data may not share the variance to such a magnitude. This is because all noise is absolutely independent and irreducible, and predicting future noise is impossible. This model could be either Over or Under fitted as the approach does not set which type of behaviour is taken, regardless fitting a model to this would risk being askewed and would result in a poorer model.

2 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

- Qualitative, these would be things that are attributes or properties that something could have or are
- Determining the integrity of a bridge in a simulator
  Predictors: Load on the bridge and materials used
  Response: Stress test results if bridge breaks or not
- Is this picture a bird or not
  Predictors: Input Photo and previous data
  Response: Yes or No based on algorithim
- Does this email look like spam
  Predictors: Email and previous data
  Response: Yes or no based on algorithim

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

- Quantatitive, these are often numerical values that something has
- Resistance of a electrical component
  Predictors: Voltage, Current
  Response: Voltage, Current
- What will the temperature be tomorrow?
  Predictors: Previous Data, Date
  Response: A temperature value
- Credit Score
  Predictors: Previous Data, Purchase History,
  Response: A Score that is carefully calculated

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

- Behavioural Clustering - Seeing how different people with different mindsets react differently to the same situation
- Determine Fruit/Veggie ripeness - Based off past data, determine how under/over/correctly ripe the input is
- Search Engine Results - Determine how close some results are than others to the user's querry

3 Problem 3

We want to learn a model to predict Y, let N denote the number of samples of data. Using calculus, derive the optimal constant fucntion under mean square error

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

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