Name: Bollempalli Leela Krishna Sai Prasanth

ID: lbs7aa

Data Mining Assignment 1

1. A) Flexible model is better because with more data, the model can train better rather than overfitting for few points

b) a flexible model would overfit the data

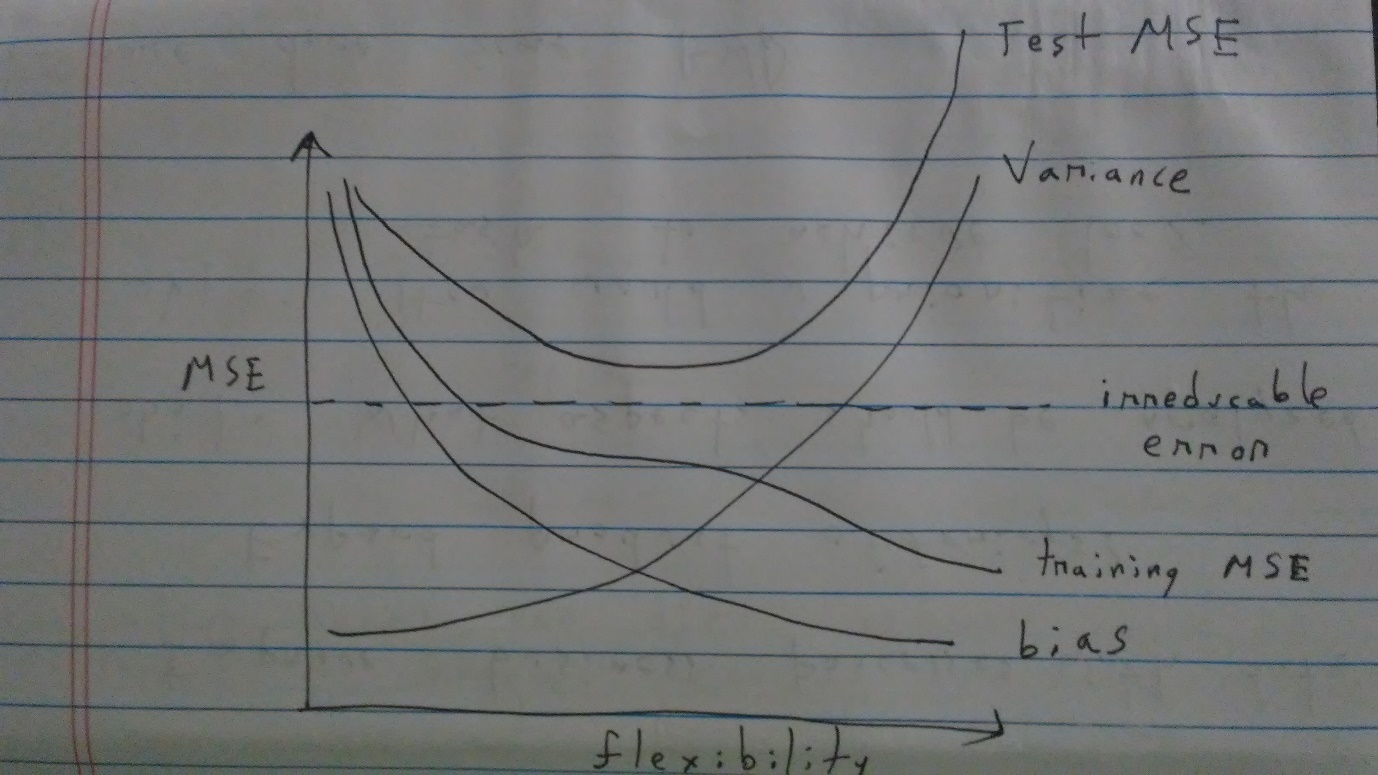
c) Flexible model is better because it can account for the necessary complexity

d) A flexible model will do worse. Flexible models will account for the noise as well

1. A) Inference. We want to know the factors

b) Prediction. Because we want to just know whether it will succeed or fail

c) Prediction. We don’t care why. We just want to know the answer

1. A) 

b) The irreducible error cannot be reduced and is fixed with flexibility

The bias will reduce considerably with a higher polynomial because higher polynomials can fit the data better

Variance of the model will increase with flexibility because the model will change drastically as we add more training data points

The training MSE will naturally decrease with a higher degree polynomial because we can fit data better

The test MSE which a mixture of variance and bias reaches a low and then starts climbing up as variance goes up

1. A) A cereal wants to know their market size in a state. They assume that people who earn $50,000 can afford cereal. Now they are trying to predict whether there is significant amount of people making more than $50K or not

Response: More than $50K income or not

Predictors: Job title, do u own a house, age, education level, etc.

Goal: Prediction

Can I use this steel bar in construction now? Is it too rusted to meet specifications? Yes or no?

Response: Can the steel bar be used? Y? N?

Predictors: Strength of steel, stress, oxygen level, humidity level etc.

Goal: Prediction

Is a customer credit worthy? Y? or N?

Response: Should a bank approve this loan application?

Predictors: Income, house rent, education level, etc.

Goal: Prediction

b) Hilton wants to understand what makes employees satisfied

response: Job satisfaction index

predictors: manager rating, tenure, age, education

Goal: Inference. I want to know which factor is having biggest factor and think about what can be done to improve job satisfaction

Under Armour wants to understand which factors have significant impact on sales

response: Under Armour sales at a store

predictors: Amount of models on display, store location, store size etc.

Goal: Inference. UA wants to understand which factors are actually affecting sales mostly

A investment focussed website is trying to predict net worth of registered users.

Response: Net worth of users

Predictors: no. of stocks checked, no. of mutual funds checked, no. of times logged in etc.

Goal: Prediction. The website can’t alter the real net worth of individuals. They are just trying to predict the net worth of the users

c) we are trying to group people into ethnic groups – height and weight

We are grouping people by spending patterns – amount spent, total items purchased

Clustering songs by theme – classic rock songs could be lengthy, blues could have less words etc.

1. A flexible approach is preferred if the actual function is high polynomial but the actual variance of the data is low. A non flexible approach is preferred if the actual function is very simple like a linear line but the variance of the actual points in high.

Also, if we are more interested in interpreting the model – understanding which predictors are affecting response variable, then a less flexible model would be preferred

1. In the case of a parametric approach, we impose a constraint on the model structure. For example, if we specify linear regression, our prediction function will generate a linear regression. However, there may be a more interesting pattern in data that we may be missing out on. Nevertheless, if we want to look at a broad trend, it might still be worth the time just to see a broad trend.

The advantage with a non parametric model is that we get to fit the data better although we may not be able to interpret the model well

1. A) table

|  |  |
| --- | --- |
| Obs | Distance |
| 1 | Sqrt(9) = 3 |
| 2 | Sqrt(4) = 2 |
| 3 | Sqrt(10) = 3.16 |
| 4 | Sqrt(5) = 2.23 |
| 5 | Sqrt(2) = 1.44 |
| 6 | Sqrt(3) = 1.73 |

b) It would be green. Since K = 1, we would look only at the closest point which is the 5th observation. The 5th observation is Green

c) we would now look at 5th, 6th, 2nd observation. The prediction would be red

d) I think K should be small. Because if we are looking at just one point, then the fit curve is going to be rough and flexible