Alexa Summers, Santhoshini Sree, Gireesh kumar Muppalla
CS 5565

Chapter 2:

1(a): Better—large training set = more flexible

1(b): Worse—small training set = less flexible

1(c): Better – relationship is non-linear = more flexible

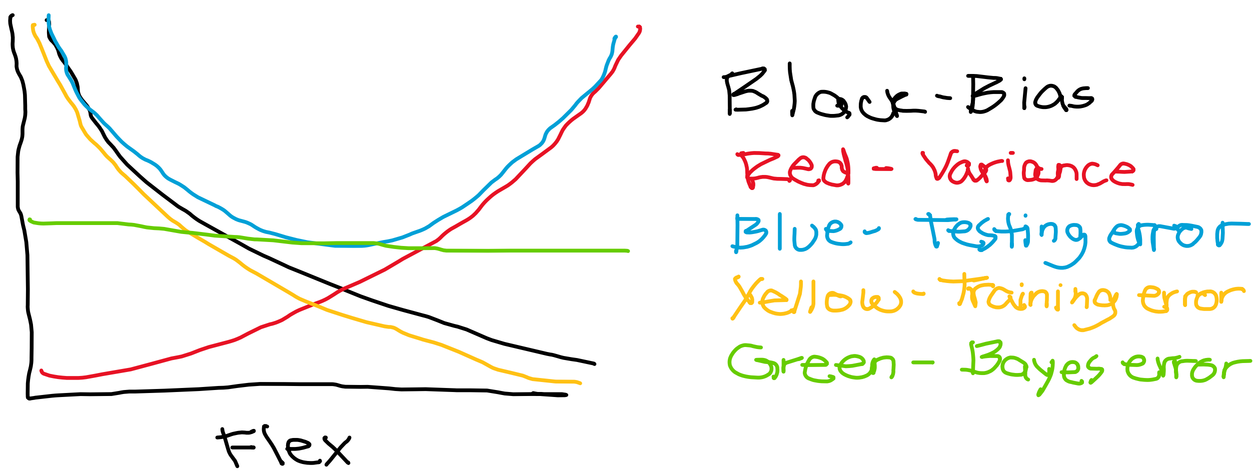
1(d): Worse— flexibility in error

2(a): Regression, Inference, N = 500, P = 3

2(b): Classification, Prediction, N = 20, P = 13

2(c): Regression, Prediction, N = 52, P = 3

3(a):



3(b): When things become more flexible or more complex, the degree of the polynomial is higher, and the degree of freedom becomes bigger. When these become bigger, the bias decreases. The error is then decreased as the flexibility grows—if there is a fit made to the training data, variance will be high. So, as flexibility becomes high, variance will start low and get higher. Training error is proportional to bias and will continue to go down with bias. Testing error starts out influenced by bias, and then later is influenced by variance. Bayes error is always constant—this does not change.

4(a): Checking email—are emails spam or not spam? This is a prediction—we are trying to guess if an email would be considered spam or not. Response: Spam/not spam. Input: sender, subject line, attachments

Assigning classes—are these courses for the bachelors or the masters curriculum? This is an inference—the inference will be based on target learning outcomes for the degrees. Response: Bachelors/masters. Input: Name, subject, difficulty, labs

Alexa Summers, Santhoshini Sree, Gireesh kumar Muppalla
CS 5565

COVID test—is the COVID test positive or negative? This is an inference—the inference will be based on current symptoms. Response: Positive/negative. Input: Temperature, Taste, Smell, Fatigue

4(b): Salary based on master agreement. Inference. Predictors: years of service, level of degree. Response: salary

Air conditioner replacement part. Inference. Predictors: age of part, intensity on part, cost of part. Response: lifespan of part

COVID death rate. Prediction. Predictors: Ventilator usage, age, lung capacity. Response: time of death after contraction

4(c): Finding groups of customers based on age and the items they purchased.

Finding groups of students based on the shape of their glasses and their degree.

Finding groups of people based on cell phone carrier and age.

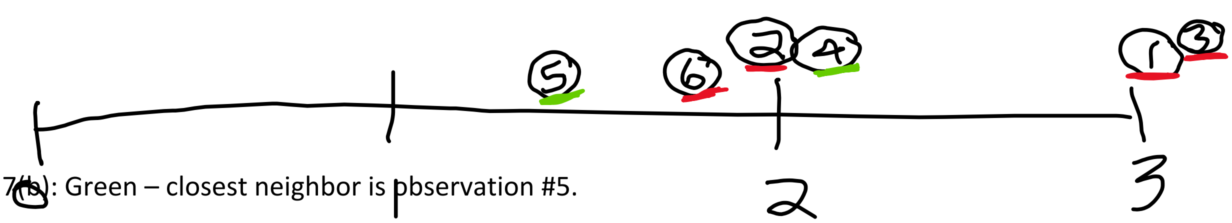
5: High flexibility is great in times of prediction because there is a low bias. High flexibility can be bad because the variance is increased and there could be an overfit. This would be most beneficial when predicting rather than trying to interpret results.

Low flexibility will be better for inference—allows interpretation of results rather than prediction.

6: In parametric, we have our estimation function of f, and with nonparametric, we just analyze without the estimation. If it's nonparametric, we can easily see something if the observation is high, however, it will be more difficult to see something if the observation is lower. In times of a lower observation, it is best to take a parametric approach because you can see some kind of linear model relationship.

7(a)

Observation	**X1**	**X2**	**X3**	**Distance**	**Y**
1	0	3	0	3	Red
2	2	0	0	2	Red
3	0	1	3	√10 = 3.2	Red
4	0	1	2	√5 = 2.2	Green
5	-1	0	1	√2 = 1.4	Green
6	1	1	1	√3 = 1.7	Red



7(b): Green – closest neighbor is observation #5.

7(c): Red – closest neighbors are 2, 5 and 6.

7(d): Small—the flexibility will be higher, so it fits better non-linearly. If the K were larger, it would fit more linearly.

Alexa Summers, Santhoshini Sree, Gireesh kumar Muppalla
CS 5565

Chapter 3:

1: The null hypothesis for TV states that TV ads have no effect on sales in the presence of radio ads and newspaper ads. The null hypothesis for Radio is that radio ads have no effect on sales in the presence of TV and newspaper ads. The null hypothesis for Newspaper is that newspaper ads have no effect on sales in the presence of TV and radio ads. The lower p-values of TV and radio signify that the hypotheses are false for TV and radio, and the higher p-value for newspaper signifies that the hypothesis is true for newspaper.

3(a): Y = 50 + 20(gpa) + 0.07(iq) + 35(gender) + 0.01(gpa * iq) - 10 (gpa * gender)

Male: 50 + 20(gpa) + 0.07(iq) + 0.01(gpa * iq) - 10 (gpa * 0)

Female: 50 + 20(gpa) + 0.07(iq) + 35 + 0.01(gpa * iq) - 10 (gpa * 1)

For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough (III).

3(b): (Gender = 1, IQ = 110, GPA = 4.0)

Female: 50 + 20(4.0) + 0.07(110) + 35 + 0.01(4.0 * 110) - 10 (4.0 * 1) = 137.1

3(c): False—Coefficients mean nothing—you need the correlation R or R^2 to provide the inference.

4(a): Because the polynomial regression can form a tighter fit against the linear data, it would most likely have a lower training residual sum of squares.

4(b): The polynomial regression would probably have a higher test RSS due to a higher error rate from overfitting (more so than linear regression).

4(c): Polynomial regression has a higher flexibility, so it will follow the points closer and reduce the training RSS.

4(d): There is not enough information because we do not know how close it is to linear. If it is closer to linear, that RSS could possibly be lower than cubic, however, if it is closer to cubic, then the cubic RSS could possibly be lower than linear.