2. Statistical Learning

1

a) Better. The model is unlikely to overfit for large dataset and small number of features.
b) Worse. Too little data for the given number of features can cause overfitting.
c) Better. We want more flexible function to capture the relationship between the predictors and response.
d) Worse. More flexible model is more likely to pick up more noise.

2

a) n=500, p=3. Regression, inference.
b) n=20, p=13, Classification, prediction.
c) n=52, p=3, Regression, prediction.

3

Bias. Decreases with flexibility until it reaches 0.
Variance. Starts higher than 0, there is always some variance. Increases with flexibility.
Irreducible error. Constant, inherent to any ML problem.
Testing error = Bias + Variance + Irreducible error.
Training error = Bias + Irreducible error - noise we capture.

When training, we fit some of the noise in the data. This makes training error lower than the sum of Bias + Irreducible error. On the other hand, this introduces variability to the estimate which is reflected in the Variance factor of the testing error.

4

a)

Application: Cancer diagnosis.
Response: Patient has cancer: yes/no.
Predictors: Values in blood sample.
Goal: Prediction.

Application: Spam classifier.
Response: Mail is spam: yes/no.
Predictors: Bag of words, sender, number of similar emails.
Goal: Prediction.

Application: IVY school acceptence.
Response: Student was accepted in IVY school: yes/no.
Predictors: High school grades, place of birth.
Goal: Inference.

b)
Application: Stock price prediction.
Response: Next week's stock price.
Predictors: Historical stock prices, news feed.
Goal: Prediction.

Application: Housing prices.
Response: Price of the house.
Predictors: Area, size, number of rooms.
Goal: Inference.

Application: Salary estimation.
Response: Expected salary.
Predictors: Age, years of education.
Goal: Inference.

c)
Cluster analysis might be helpful for:
- Market segmentation for better customer targeting.
- Anomaly detection, such as broken engines.

5

Very flexible methods are generally preferred to the less flexible ones because of the higher predictive power.

There are two scenarios where more flexibility might not be desirable: - We don't have enough training data for the chosen number of features. - We want our model to be more interpretable. E.g. we might prefer linear regression.

6

Parametric approach has pros and cons.

Pros:
- Interpretable. We can investigate relationships between features and output.
- Clearly defined. We define the predictive function. This is simpler than estimating arbitrary function.
- Needs less data. If the predictive function is simple, we don't need much data to fit.

Cons:
- Less flexible. We need many parameters to capture highly non-linear relationship between features and output.

7

a)

Obst	Dist	Y
1	3	R
2	2	R
3	$\sqrt{10}$	R
4	$\sqrt{5}$	G
5	$\sqrt{2}$	G
6	$\sqrt{3}$	R

b) Green. Observation 5 which is the closest point is green.

c) Red. Closes three observations are 2, 5, and 6. Two of which are green.

d) K will be small. If one wants to capture the decision boundary of highly non-linear function, the function should be able to quickly adjust to the local changes. This is only possible with small K.