1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
   1. The sample size n is extremely large, and the number of predictors is small.

* Answer: A flexible model would do better because there is more data, and therefore, harder for the model to overfit the data
  1. The number of predictors p is large and the number of observations n is small
* Answer: A non flexible model would do better because there is less data, and it would be too easy to overfit the training set
  1. The relationship between the predictors and response is highly non-linear
* Answer: A flexible model would do better, because a linear model would not accurately estimate the true function, f(x)
  1. The variance of the error terms, i.e. 𝜎2 = Var(∊), is extremely high
* Answer: Not enough information because this is the irreducible error, and neither model will be able to improve on this type of error

1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally , provide n and p.
   1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO Salary.

* Answer: Regression problem, focusing on inference, with 500 observations (n), and 4 parameters (p).
  1. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
* Answer: Classification problem, focusing on prediction, with 20 observations (n), and 13 parameters (p).
  1. We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
* Answer: Regression problem, focusing on prediction, with 52 observations (n), and parameters (p).

1. We now revisit the bias-variance decomposition.
   1. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



* 1. Explain why each of the five curves has the shape displayed in part (a).
* Answer:
  + Bayes - flat line because it is the irreducible error. Will stay the same regardless of model flexibility. (Model cannot reduce the irreducible error)
  + Test Error - starts high, decreases as model is more flexible (reduces error), but then increases at end due to overfitting the training data
  + Train Error - decreases and approaches bayes error as the model overfits
  + Bias - Is the error from calculating complex real life models with a simple model, so inherently, as the model becomes more complex, the bias will lower.
  + Variance - as the data is overfit, small changes in the dataset will cause large changes in the model accuracy, therefore, as flexibility increases, variance increases as well

1. You will now think of some real-life applications for statistical learning.
   1. Describe a real-life application in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

* Answer: Insurance companies might want to classify a newly written auto policy as prone to lapse or not. If the policy is prone to lapse, the company may charge a higher rate to help offset the high cost of writing an auto policy for a short time period. The response (Y) is lapsing or not at the end of the policy term. Predictors could be the applicants age, gender, and credit score. The goal would be both inference and prediction. Inference will help the company target certain customer demographics that tend not to lapse. Prediction will help offset costs.
  1. Describe a real-life application in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
* Answer: Predicting sales of a new store branch based on certain demographics, grabbed from existing branches. The response (Y) would be total sales, and predictors could be median household income of the zip code(s), median head of household age, and most popular car brand of area. The goal would mostly focus on prediction accuracy, however, some inference may be useful for future business growth/targeting.
  1. Describe a real-life application in which cluster analysis might be useful.
* Answer: MLB teams can try to group players together based on their average, on base percentage, home runs, and other statistics, in order to more accurately project their future growth prospect.

1. What are the advantages and disadvantages of a very flexible approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

* Answer: A more flexible approach tends to optimize the bias-variance trade off, as well as better estimates the distribution (f). A less flexible approach may be preferred if the distribution is known to be simple, or when inference is the main goal.

1. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification? What are its disadvantages?

* Answer: Parametric approaches make an assumption about the distribution of f. These models tend to be simple, and their main advantage is interpretability and less of a requirement for a large number of observations. Their disadvantage is assuming a distribution/simplifying a problem that is likely much more complex.

1. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y |
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | -1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for Y when X1=X2=X3=0 using K-nearest neighbors.

1. Compute the Euclidean distance between each observation and the test point, X1=X2=X3=0

|  |
| --- |
| Euclidean Distance |
| 3 |
| 2 |
| 3.162 |
| 2.236 |
| 1.414 |
| 1.732 |

1. What is our prediction with K=1? Why?

* Answer: Green, because observation 5 was closest to the test point

1. What is our prediction with K=3? Why?

* Answer: Observations 2, 5, and 6 are the closest to our test point. 2 of the 3 points are Red, so our prediction is red

1. If the Bayes decision boundary in this problem is highly non linear, then would we expect the best value for K to be large or small? Why?

* Answer: We would expect the best value for K to be small, because as K increases, the decision boundary line approaches linear.