**CHAPTER 2: STATISTICAL LEARNING**

**Conceptual:**

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 p is small.

Answer: When a given sample is extremely large and contains a small number of predictors, a flexible approach will perform better than an inflexible approach.

1. The number of predictors p is extremely large, and the number of observations n is small.

Answer: When the number of predictors, p, is extremely large and the number of observations, n, is small, then a flexible approach would perform worse than an inflexible approach because a flexible approach will overfit the data.

1. The relationship between the predictors and response is highly non-linear.

Answer: When the relationship is non-parametric, then a flexible model would perform better simply because that is what it is geared to do. The relationship is *highly* non- linear.

1. The variance of the error terms, i.e. σ 2= Var(ε), is extremely high.

Answer: If the variance of the error terms is extremely high, then we should expect a flexible approach to perform better because it maps closer to the observations.

2) 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: n=500 and p=3. This is a regression and inference problem since we are interested in understanding which factors affect CEO salary.

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: n=20 and p=13. This is a classification problem since we are dealing with a binary outcome (success or failure) and prediction since we are only interested in the outcome.

1. We are interested in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Answer: n=52 and p=3. This is a regression and prediction problem since we are interested in predicting the % change in the US dollar.

3) 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.

Answer:

A close up of a device

Description automatically generated

1. Explain why each of the ﬁve curves has the shape displayed in part (a).

Answer:

* **Irreducible Error**: is a constant random and is irrespective of model flexibility because the noise is random.
* **Training Error**: decreases as model flexibility increases because more flexible models tend to fit the noise more closely in a dataset.
* **Testing Error**: usually decreases as model flexibility increases but at a certain point it flattens out and starts to increase again. This is due to the fundamental properties of flexibility.
* **Variance:** increases in parallel with model flexibility because as a model becomes more flexible, its variability also increases.
* **Squared Bias:** bias decreases as flexibility increases because as we increase flexibility the model maps closer to the residuals (hence the bias variance trade-off).

4) You will now think of some real-life applications for statistical learning.

1. Describe three real-life applications in which classiﬁcation might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction?

Answer:

1) *Response:* Success or failure of a COVID-19 vaccine. *Predictors:* health history, current medications, virus mutation, etc. *Goal:* Prediction.

2) *Response:* Success of a movie. *Predictors:* Cast, cast skills, budget, synopsis, viewer age, director vocabulary, cast skills, setting, length, songs, polarity, genre, and so on. *Goal:* Prediction.

3) *Response:* accepting a candidate for a job. *Predictors:* skills, passion, network, experience, punctuality, commitment, ability to think/act quickly and deliver results, etc. *Goal:* Prediction.

1. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction?

Answer:

1) *Response*: Unemployment rate in U.S.A. due to COVID-19 during 2020. *Predictors*: positive COVID-19 cases, COVID-19 cases that have been cured, rate of growth (+/-) for COVID-19, geographical spread, and so on. *Goal*: Inference

2) *Response:* Housing prices rate increase/decrease in U.S.A. due to corona virus. *Predictors:* GDP increase/decrease, housing demand, unemployment, market indexes, currency valuation, safety, crime, structure of home, positive COVID-19 cases, COVID-19 cases that have been cured, rate of growth (+/-) for COVID-19, geographical spread, and so on. *Goal:* Inference.

3) *Response:* LIRR and MTA fare hike %. *Predictors:* loss due to COVID-19, funding, repairs required, season, basic supply/demand rates, traffic, delivery time, and so on. *Goal:* Inference. *Goal:* Inference.

1. Describe three real-life applications in which cluster analysis might be useful.

Answer:

1) *Response:* Cancer type diagnostic. *Predictors:* cat scans, synergistic health conditions, side effects of medication, age, and so on. *Goal:* Prediction.

2) *Response:* which med school a student will attend. *Predictors:* GPA, extracurricular activities, publications, research, sports, social class, ethnicity, essay quality, field of interest, scholarship, commute time, cost, distance from loved ones, and so on. *Goal:* Prediction.

3) *Response:* Low, medium, or high success of online articles. *Predictors:* numeric word characteristics polarity, positive, negative, etc. *Goal:* Prediction.

5) What are the advantages and disadvantages of a very ﬂexible (versus a less 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:

* Advantages: a very flexible approach may provide a better fit for non-parametric data and decreases the bias.
* Disadvantages: a very flexible approach may *overfit* the noise in the data (irreducible error) and consequently, considerably increase the error in testing data.
* We may consider a more flexible approach than a less flexible approach when the purpose of the model is to predict a certain outcome, stock price range for instance.
* We may consider a less flexible approach when the purpose of the model is interpretability as less flexible models are (generally) more easily interpreted than flexible models.

6) 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 (as opposed to a nonparametric approach)? What are its disadvantages?

Answer: Similar to 5 above.

7) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

A picture containing clock

Description automatically generated

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation | X1 | X2 | X3 | Y | Distance |
| 1 | 0 | 3 | 0 | Red | 3 |
| 2 | 2 | 0 | 0 | Red | 2 |
| 3 | 0 | 1 | 3 | Red | 3.16 |
| 4 | 0 | 1 | 2 | Green | 2.23 |
| 5 | -1 | 0 | 1 | Green | 1.41 |
| 6 | 1 | 1 | 1 | Red | 1.73 |

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

If K=1, then K is closest to a neighbor to a distance of 1.41, or our 5th observation x5. Therefore, Green.

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

If K=3, then K is closest to X2 (red), X5,(green)and X6 (red).

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

Small because as K increases, the Bayes decision boundary becomes more and more linear.