2.4 Exercises

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.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The relationship between the predictors and response is highly non-linear.

   (d) The variance of the error terms, i.e. $\sigma^2 = \mathrm{Var}(\epsilon)$, is extremely high.

   Solution.

   (a) A flexible method will be better in this scenario since there are many data points. An inflexible method would not be able to capture the trend with many data points well. However, one must keep in mind that with a large sample size, the flexible method will track more sensitively with the noise in the data, which will contribute to error.

   (b) If the number of observations n is small, then an inflexible method is better. A very flexible method will overfit the data and may not be able to visualize a trend well.

   (c) If the response is highly non-linear, a flexible method is better, as it will be able to fit this non-linear response better than an inflexible method.

   (d) If the variance of error terms is extremely high, then an inflexible method is better, as it will be less affected by noise than a highly flexible method.

   □

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.

   (a) 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.

   (b) 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.

   (c) 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.

Solution.

(a) This is a regression problem, and we are interested in inference. Here, $n = 500$ and $p = 4$.

(b) This is a classification problem, and we are interested in prediction. Here, $n = 20$ and $p = 14$.

(c) This is a regression problem, and we are interested in prediction. Here, $n = 52$ and $p = 4$.

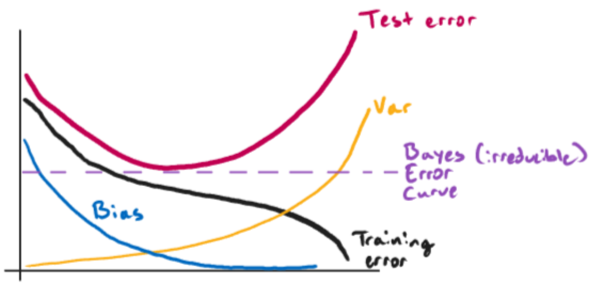
□

3. We now revisit the bias-variance decomposition.

   (a) 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.

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

Solution.



(a)

(b) As flexibility starts to increase, the shape of the data is better modeled and the test error decreases towards the optimal solution. As the solution curve better fits the data, bias decreases. However, as the solution curves deviates from a linear solution to meet each point, variance begins to increase. Overall, as flexibility increases, training error monotonically decreases. However, at a certain point, overfitting occurs, so while bias goes to zero, the solution curve fits with the noise, and variance increases as does test error.

□

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

   (a) 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? Explain your answer.

   (b) 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? Explain your answer.

   (c) Describe three real-life applications in which cluster analysis might be useful.

Solution.

(a)

   (1) **Determining bird species from chirping sound profile.** Predictors: frequency response of chirp sounds recorded in the wild. Response: Bird species from known data about bird chirp sound pitches. Goal is inference.

   (2) **Lie detector.** (Watching true crime while writing this) Predictors: Heart rate, movement, blood pressure, steady eye contact (deviation from axis connecting suspect's eyes to interrogator's). Response: Categorizing a suspect's statement as a lie or truth. The goal is inference.

   (3) **Political ideology.** Predictors: Answers on a quiz of political questions asking the participant their view. Response: Categorizing the respondent's political ideology. The goal is prediction.

(b)

   (a) **Mile time.** Predictors: Age, height, weight, body fat percentage. Response: Mile time. The goal is to predict someone's mile time.

   (b) **Course performance.** Predictors: Homework grades, quiz grades, exam scores. Response: Overall course grade. The goal is inference, and to determine if a certain aspect of the syllabus has a larger-than-intended impact on the course grade.

   (c) **Solid-State Drive Lifespan.** Predictors: Read/writes per day. 1 year read/write speed trends. Response: Solid-State drive lifespan. The goal is to predict how long a SSD will last before failure.

(c)

   (a) **Turbulence modeling.** Consider particles injected into a stream of fluid captured on a high speed camera. Clusters of particles indicate flow patterns and vortices.

   (b) **Penguin groups.** Aerial footage of penguins waddling around to identify families/tribes/similar species.

   (c) **Sock distribution.** Distribution of socks around my apartment, used to infer which rooms I tend to occupy the most.

□

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?

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?

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

Obs.	X_1	X_2	X_3	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 $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) 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?

Solution.

(a) We have,

Obs.	1	2	3	4	5	6
Dist.	3	2	3.16	2.24	1.41	1.73

(b) For $K = 1$, the closest observation is #5, which is classified as green. And so, for $K = 1$, our prediction is green.

(c) For $K = 3$, the closest three observations are #5, #6, and #4. Since 2/3 (the majority) of these observations are green, our prediction is green.

(d) Since the Bayes decision boundary in this problem is highly nonlinear, we expect the best value for K will be small. As K is large, the decision boundary will not be sufficiently flexible.

□

Applied

8. This exercise relates to the *College* data set, which contains a number of variables for 777 different universities and colleges in the US. The variables are:

   - Private: Public/private indicator
   - Apps: Number of applications received
   - Accept: Number of applicants accepted
   - Enroll: Number of new students enrolled
   - Top10perc: New students from top 10% of high school class
   - Top25perc: New students from top 25% of high school class
   - F.Undergrad: Number of full-time undergraduates
   - P.Undergrad: Number of part-time undergraduates
   - Outstate: Out-of-state tuition
   - Room.Board: Room and board costs
   - Books: Estimated book costs
   - Personal: Estimated personal spending
   - PhD: Percent of faculty with Ph.D.s
   - Terminal: Percent of faculty with terminal degree
   - S.F.Ratio: Student/faculty ratio
   - perc.alumni: Percent of alumni who donate
   - Expend: Instructional expenditure per student
   - Grad.Rate: Graduation rate

   Before reading the data into Python, it can be viewed in Excel or a text editor.

   (a) Use the `pd.read_csv()` function to read the data into Python. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

   (b) Look at the data used in the notebook by creating and running a new cell with just the code `college` in it. You should notice that the first column is just the name of each university in a column named something like `Unnamed: 0`. We don't really want pandas to treat this as data. However, it may be handy to have these names for later. Try the following commands and similarly look at the resulting data frames:

   ```
   college2 = pd.read_csv('College.csv', index_col=0)
   college3 = college.rename({'Unnamed: 0': 'College'}, axis=1)
   college3 = college3.set_index('College')
   ```

   This has used the first column in the file as an index for the data frame. This means that pandas has given each row a name corresponding to the appropriate university. Now you should see that the first data column is `Private`. Note that the names of the colleges appear on the left of the table. We also introduced a new python object above: a dictionary, which is specified by (key, value) pairs. Keep your modified version of the data with the following:

```
college = college3
```

(c) Use the `describe()` method to produce a numerical summary of the variables in the data set.

(d) Use the `pd.plotting.scatter_matrix()` function to produce a scatterplot matrix of the first columns `[Top10perc, Apps, Enroll]`. Recall that you can reference a list `C` of columns of a data frame `A` using `A[C]`.

(e) Use the `boxplot()` method of `college` to produce side-by-side boxplots of `Outstate` versus `Private`.

(f) Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
college['Elite'] = pd.cut(college['Top10perc'],
                                            [0, 0.5, 1],
                                  labels=['No', 'Yes'])
```

Use the `value_counts()` method of `college['Elite']` to see how many elite universities there are. Finally, use the `boxplot()` method again to produce side-by-side boxplots of `Outstate` versus `Elite`.

(g) Use the `plot.hist()` method of `college` to produce some histograms with differing numbers of bins for a few of the quantitative variables. The command `plt.subplots(2,2)` may be useful: it will divide the plot window into four regions so that four plots can be made simultaneously. By changing the arguments you can divide the screen up in other combinations.

(h) Continue exploring the data, and provide a brief summary of what you discover.

9. This exercise involves the *Auto* data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the `min()` and `max()` methods in numpy.

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

10. This exercise involves the *Boston* housing data set.

    (a) To begin, load in the *Boston* data set, which is part of the `ISLP` library.

    (b) How many rows are in this data set? How many columns? What do the rows and columns represent?

    (c) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

    (d) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

    (e) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

    (f) How many of the suburbs in this data set bound the Charles river?

    (g) What is the median pupil-teacher ratio among the towns in this data set?

    (h) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

    (i) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.