Ch 2. Statistical Learning

Fraser Watt

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical 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.

A flexible model would perform better in this scenario. The large dataset means our model can capture more complex relationships between the predictors and response, identifying patterns in the data that a less flexible might not be able to.

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

We would expect a more flexible method to perform worse in this scenario. Flexible models generally require much more data to train properly, and this problem would be exacerbated by a high dimensional space. The model is likely to latch onto spurious relationships in the data without a large n to train on. This method is more likely to overfit than a parametric method would be.

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

We would expect a flexible method to fit the data better in this scenario. Flexible methods are better equipped for finding non-linear relationships between the response and its predictors, as they do not make assumptions about the shape of f.

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

In this scenario, our flexible method would be expected to perform poorly. The error is made up of both reducible (variables we could have used as predictors) and irreducible error (unobservable factors, events out of our control, or noise), and a flexible method applied to a dataset with high variance in the error term would be at risk of training on spurious relationships in the irreducible error.

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.

This is a inferencial regression problem. n is the top 500 firms in the US, and p is record profit, number of employees, industry and CEO salary.

b. We are considering lauching 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.

This is a predictive classification problem. n is the 20 similar products, whilst p is whether it was a success or failure, price charged for the product, marketing budget, competition price, and these "ten other variables".

c. We are interesting 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.

This is a predictive regression problem. n is the weekly stock data for 2012. p is the % change in the dollar, the % change in the US market, the % change in the British Market, and the % change in the German market.

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).

GO BACK TO THIS WHEN YOU HAVE A TABLET TO HAND.

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.

- Categorising high value leads in a sales pipeline. This would be a prediction task, where the response would be a sales category ('SME', 'Enterprise Customer', etc), and predictors could be number of employees, any publicly available revenue, social media followers, country of headquarters.
- Recognising benign vs malignant tumors as part of an image processing model. This would be a prediction problem, where the response would be a binomial label of 'Benign' or 'Maligant' depending on the type of tumor. The predictors would be each individual pixel in the image being scanned.
- Flagging hate speech on a social media platform. This would be a prediction model, where a binomial 'Hate Speech' / 'Not Hate Speech' classifier would be trained on a corupus of tweets on the social media platform. Key words which are deemed to add or subtract the probability of hate speech would act as predictors in one way or another.

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.

- Predicting stock prices. This would be a prediction task, where you would be predicting the movement of a given stock in a market across unseen time series data. Predictors would be past performance, and you would probably also want to enrich the model with additional data such as Google News alerts, or sentiment analysis of social media posts.
- Understanding the relationship between revenue and marketing spend on different ecommerce traffic sources. This would be an inferencial task, where the predictor would be revenue, and the predictors would be TV spend, radio spend, Google Adwords spend, Facebook Ads spend, date, etc.
- Capacity planning for a hospital, so that you knew how many hospital staff to assign to a given shift. The response could be patients in the ward, and predictors could be the number of patients in ward year to date (and averaged over the last 30 days), temperature, local population.

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

- Grouping similar articles together for a news aggregation service who were thinking about rebuilding their article tag hierarchy.
- You might want to perform cluster analysis on a web store's customer behaviour in order to formulate ideas about different clusters of customers.
- A molecular biologist might want to analyse gene expression by looking at similar groups of genes, which could be clustered together using statistical learning techniques.

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?

The advantages of a very flexible approach is that it can model more complex (especially non-linear) relationships, as more flexible approaches make fewer assumptions about the shape of the function f. The disadvantages are that more flexible approaches require more data to train accurately, are difficult to interpret. They have the potential to out-perform less flexible methods, although this reduced bias makes them more susceptable to overfitting.

A more flexible approach might be preferred when you have a lot of data and are dealing with a pure prediction problem such as automated stock trading, as we are less concerned with visualising the model or interpreting the results. Conversely, if we are using statistical learning to better understand the relationship between the predictors and response, we may be better off using a less flexible (and therefore more interpretable) model like linear regression or a decision tree. More flexible methods are also useful when you have confirmed that you are working with a non-linear relationship between predictors and response.

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 non-parametric approach)? What are its disavantages?

Parametric models make assumptions about the structure of the function (f) you require to fit your model to the data, and then perform an optimisation function on that method in order to create a fitted model. On the other hand, non-parametric models make far fewer assumptions about the "shape" of f.

The advantages of a parametric model are that it is easier to explain, especially to non-technical stakeholders, and can be done with significantly less data and computational power. The disadvantages are that the structure we assume the function f takes is realistically not going to be the same as the "true" shape of f, which will limit its predictive power.

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

```
knn_df <- data.frame(
    X1 = c(0, 2, 0, 0, -1, 1),
    X2 = c(3, 0, 1, 1, 0, 1),
    X3 = c(0, 0, 3, 2, 1, 1),
    Y = c("Red", "Red", "Red", "Green", "Green", "Red")
)
knn_df
```

```
##   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
```

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

- i_1 **and test point:** $\sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = 3.0000$
- i_2 **and test point:** $\sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = 2.0000$
- i_3 **and test point:** $\sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} \approx 3.1623$
- i_4 **and test point:** $\sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} \approx 2.2361$
- i_5 **and test point:** $\sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} \approx 1.4142$
- i_6 **and test point:** $\sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} \approx 1.7321$

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

We would predict 'Green', as the closest single point is i_5 and $k = 1$.

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

We would predict 'Red', as the closest three samples are i_5 ('Green'), i_6 ('Red') and i_2 ('Red'). Majority vote wins, so our test point would be classified as 'Red'.

d. 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?

In the scenario where the Bayes decision boundary is non-linear, we would expect a smaller K to outperform a larger one. When the "true f" takes on a complex shape, a larger value for K will tend to "smooth" out highly non-linear decision boundaries giving us a simpler model. Where we have a highly non-linear relationship between predictors and response, we will want a smaller K to create a more complex boundary and disregard points further away from the unseen data.