ISLR-R21._1

Mücahit Zor

2023-08-23

Table of contents

Preface

Here are my notes on *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013*

1 What Is Statistical Learning?

Question: How to improve sales of our product?

We have a data set:

```r
advertising = read_csv("./data/Advertising.csv") %>% as_tibble %>% select(-1)
```

```
New names:
Rows: 200 Columns: 5
-- Column specification
-------------------------------------------------------- Delimiter: "," dbl
(5): ...1, TV, radio, newspaper, sales
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
```

```r
advertising
```

```
# A tibble: 200 x 4
      TV radio newspaper sales
   <dbl> <dbl>     <dbl> <dbl>
 1 230.   37.8      69.2  22.1
 2  44.5  39.3      45.1  10.4
 3  17.2  45.9      69.3   9.3
 4 152.   41.3      58.5  18.5
 5 181.   10.8      58.4  12.9
 6   8.7  48.9      75     7.2
 7  57.5  32.8      23.5  11.8
 8 120.   19.6      11.6  13.2
 9   8.6   2.1       1     4.8
10 200.    2.6      21.2  10.6
# i 190 more rows
```

$n = 200$, independent variables (predictors) are `TV`, `radio`, and `newspaper` advertising spendings in thousands of dollars. We want to explore their relationship with `sales`; quantity of product sold for each advertising mixture. If we determine association between advertising and sales, we can provide adjustment of advertisement budgeds based on most effective media to increase sales; we want to develop an accurate model that ca be used to predict sales on the basis of three media budgets.

We denote all input variables (actual–realized) as $X_1, X_2, ..., X_p$ and use X to refer all of them. In this case $X = (X_1, X_2, X_3)$. Sales is denoted with Y.

This means we assume a relationship between Y and X in a form of

$$Y = f(X) + \epsilon$$

(2.1)

- Here f is some fixed, but unknown function of X.
- ϵ is a random *error term* => independent of X and has a mean zero.

So f represents systematic information that X provides about Y.

f is generally unknown. We will need to estimate f baed on the observed points => $\hat{f}$.

Statistical learning refers to a set of approaches for estimating f.

1.1 Why estimate f?

Two reasons: * *prediction* * *inference*

Prediction

Most of the time we have X but we might not have Y. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}X$$

(2.2)

Here $\hat{f}$ is treated as a *black box*. We are not concerned with the exact form of $\hat{f}$, we just want to have accurate predictions of Y.

Imagine we have $X = (X_1, X_2, \dots, X_p)$; blood sample characteristics of patients. Y is a variable showing the patient's risk for a adverse reaction to a drug. We don't want to give the drug and see the reaction, so we want to predict reactions.

The accuracy of our predictions $\hat{Y}$ of Y, depends on two quantities:

- *reducible error*

  Generally $\hat{f}$ will not be a perfect estimate for f. This inaccuracy will introduce some error, which we call reducible error since we can improve our accuracy of $\hat{f}$ using the most appropriate statistical leraning method.

- *irreducible error*

  Even if we estimate f perfectly, our estimated response would take the form $\hat{Y} = f(X)$; our predictions would still get some error. This is because Y is not just a function of X but also a function of ϵ, which cannot be predicted by X. So the level of ϵ would also effect our prediciton accuracy. And we cannot remove this error; thus, irreducible.

  ϵ is larger than zero; because ϵ may contain some variables we don't include in our model, but effect Y.

Inference

Here we want to understand the way that Y is affected by X. In this setting, we wish to estimate f but we are not concerned with predicting. We want to understand the relationship between X and Y; how Y changes as X changes. We **don't** treate $\hat{f}$ as a *black box* now since we need to know its exact form. In this setting we are interested in answering questions such as

- *Which predictors are associated with the response?*

  Usually not all predictors are associated with Y. We need to identify the *important* predictors among a large set of possible predictors.

- *What is the relationship between the response and each predictor?*

  Some predictors have positive some negative association with Y. Depending on the complexity of f, the relationship between Y and X_i may also depend on the values of other predictors(X_j) => *synergy*

- *Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?*

Sometimes we are interested with prediction: Identifying individuals who will respond positively to a mailing, based on observations of demographic variables. Here we are not interested with understanding the relationship of demographic variables and response, we just want an accurate model to predict the response using the predictors. This is prediction.

But often we are interested to answer questions like: *Which media contribute to sales?*, *Which media generate the biggest boost in sales?*, or *How much increase in sales is associated with a given increase in* `TV` *advertising?*. This is inference.

And sometimes we want a combination of both: *Values of homes based on crime rate, zoning, distance from a river, air quality, schools, size of houses etc.* and *How does air quality effect valeus of homes?.*

We use different models for prediction, inference, or combination of the two.

1.1.1 How Do We Estimate f?

There are many linear and non-linaer approaches we will discuss. But generally these models share certain characteristics. Here are they:

- We will always assume that we have observed a set of n different data points. These data points, observations, are called *training data*; which we will use these observations to train, or teach, our model on how to estimate f. Our training data will consist of $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

We want to apply a statistical learning method to the training data to estimate the unknown function f. We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any obsrvation (X, Y).

These statistical learning methods can be charactarized as either *parametric* or *non-parametric.*

Parametric Methods

Parametric methods involve a two step model-based approach:

1. *Select a model => Make an assumption about the functional form of f*: is it linear, non linear?

For example a linear f assumption would yield a *linear model*

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

(2.4)

2. *Fit* or *train* the model

After we select a model, we need a procedure that uses training data to *fit* or *train* the model.

For linear model, we need to estimate the parameters of the model $(\beta_0, \beta_1, \dots, \beta_p)$. That is we want to find values of these parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

The most common approach to fitting the model (2.4) is called *ordinary least squares.* Chapter3. But there are other approaches as well.

This model-based approach is called *parametric:* we estimate f via estimating a set of parameters.

Disadvantage (potential): model we choose will usually not match the true unknown form of f => our estimates will be poor. => solution: choose a *flexible* model that can fit different possible functional forms for f => you will need to estimate more parametrs => *overfitting the data.*

Non-parametric Methods

=> No explicit assumptions about the functional form of f. The goal is to get an estimate of f that gets as close to the data points as possible without being too rough or wiggly => advantage over parametric approach: no assumption about the functional form of f–potentially accurately fit a wider range of possible shapes for f.

Disadvantage => lots of parameters to estimate => very large of observations required to obtain an accurate estimate for f.

1.1.2 The Trade-Off Between Prediction Accuracy and Model Interpretability

Some models are flexible some restrictive; in the sense that they can produce just a small range of functional forms to estimate f. Linear regression for instance is a relatively inflexible approach. Other metgods such as thin plate splines (non-parametric) are more flexible because they can generate a much wider range of possible functional forms to estimate f.

Why would be ever choose to use a more restrictive method instead of a very flexible approach? :

- If we are mainly interested in inference, restrictive models are more interpretable. They give more information about each predictors effect on predicted.
- If we are mainly interested in prediction, flexible models give better fit. => but may yield less accurate fits due to *overfitting!*

1.1.3 Supervised vs Unsupervised Learning

Most statistical learning problems fall into these two categories: *supervised* or *unsupervised.*

In supervised learning for each observation of the predictor values $x_i, i = 1, \dots, n$ there is an associated response value y_i. We wish to fit a model that relates the response to the predictors with the aims of either accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). Linear regression, GAM, boosting, support vector machines operate in the supervised learning domain.

Unsupervised leraning describes a situtaion in which for every observation $i = 1, \dots, n$ we obser a vector of values x_i but no associated response y_i. We cannot use a linear regression model since we dont have y_i values. Here we can seek to understand the relationships between the variables or between the observations; like *cluster analysis*, or clustering: to assert on the basis of $x_i, \dots, x_n$ whether the observations fall into relatively distinct groups.

1.2 Regression vs Classisfication Problems

Variables can be characterized as either *quantitative* or *qualitative*(also known as *categorical*). Quantitative varaibles take on numberical values: a person's age, height, or income, the value of a house, price of stock. Qualitative varaibles take on values n one of K different *classes*, or categories: aperson's gender(male or female), the brand of a good (A,B, or C), a person's race etc.

We refer to problems with a quantitative response as *regression* problems, and probmes with a qualitative response as *classification* problems. However, the distinction is not clear-cut.

Least squares regression is used with a quantitative response, whereas logistic regression is typically used with a qualitative response. Some statistical methods, such as K-nearest negihbors and boosting, can be used in the case of either quantitative or qualitative.

We usually select statistical learning methods based on whether the response is quantitative or qualitative: we might use linear regression wjen quantitative and logistic regression when qualitative. But whether the *predictors* are qualitative or quantitative is usually not that important. Most of the statistical learning methods can be applied regardless of the predictor varible type.

1.3 Assessing Model Accuracy

There is no one method that dominates all others over all possible data sets. On a particular data set, one metghod may work best, but some other method may work better on a similar but differet data set. So it is important to assess the model accuracies of the methods.

Here some ways to asses the model accuracy

1.3.1 Measuring the Quality of Fit

So, to evaluate the performance of a statistical learning method on a given data set, we need to measure how well its predictions actually match the observed data.

In the regression setting, the most commonly-used measure is the *mean squared error (MSE)*, given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

MSE will be small if the predicted responses are veryt close to the true responses, and large if predicted and true responses differ substantially **on average**.

Here since MSE is computed using the training data it is best to refer it as **training MSE**. But in general, we do not really care how well the method works on the training data => *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.*

Imagine: stock price prediction => we have training and test data => we already know the stock prices of the past, we dont care about the training data accuracy of the model, we want our model to predict the future prices of stocks best.

Or we have blood characteristics of diabetes patients. We don't want our model to explain our existing patient's classification of diabetes or not, we want our model to predict our future patien's situation the best.

Mathematically:

We fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and we obtain the estimate $\hat{f}$. We can then compute $\hat{f}(x_1), \ldots, \hat{f}(x_n)$. If these are approximately equal to $y_1, \ldots, y_n$ then our training MSE will be small. Howeer, we are not interested in whether $y_i \approx \hat{f}(x_i)$, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a *previously unseen test obsrevation not used to train the statistical learning method.*

That is, we want to choose the method that gives the lowest *test MSE*!.

So with our *test data* we can compute *test MSE*

$$\text{MSE}_{test} = \frac{1}{n_{test}}\sum(y_{test_i} - \hat{f}(X_{test_i}))$$

(2.6)

We want the test MSE to be small as possible. We can compute test MSE via (2.6) if we have test data for different models and select the model with minimum test MSE.

If we don't have a test data, you might think our goal would be to minimize the training MSE since test and training data are colesly related. But no; minimal training MSE doesn't guarantee minimal training MSE

Usually as the level of flexibility increases, the curves fit the observed data more closely => lower training MSE. The level of flexibility is quantified by *degrees of freedom*. More restricted models have lower degrees of freedom. and usually the training MSE declines as flexibility increases.

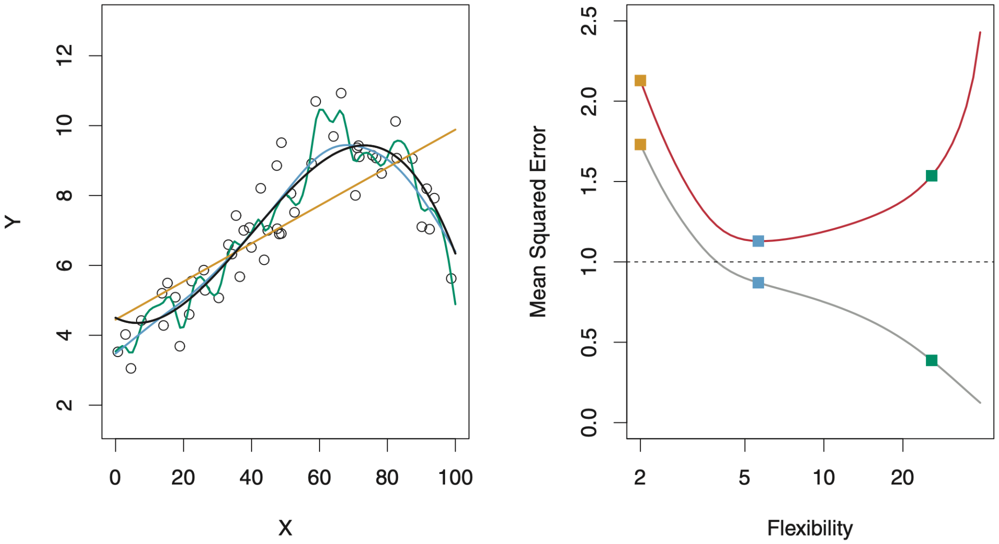

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

As the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a *U-shape* in the test MSE. This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of the statistical method being used. AS model flexibility increases, training MSE will decrease, but the test MSE may not. When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data. This happens because our statistical learninig procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random change rather than by true properties of the unknown function f. When we overfit the trainin data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don't exist in the test data.

Note that regardless of whether or not overfitting has occured, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or inderectly seek to mimizie the training MSE. Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

In practice, training MSE is computed easily, but estimating test MSE is hard because usually no test data are available. We will learn approaches that can be used in practice to estimate the mininmum test MSE. One important method is *cross-validation*(Chapter 5), which is a method for estimating test MSE using the training data.

1.3.2 The Bias-Variance Trade-Off

The U-shape in the test MSE result of two competing properties of statistical learnig methods. The expected test MSE, for a given value x_0 can always be decomposed into sum of *variance* of $\hat{f}(x_0)$, the squared *bias* of $\hat{f}(x_0)$ and the variance of the error terms ϵ. That is

This means that to minimize the expected test error, we need to simultaneously have *low variance* and *low bias*. Since variance is always bigger than zero; $\mathrm{Var}(\epsilon)$, and $\mathrm{Bias}(\hat{f}(x_0))$ are nonnegative. So, the expected test MSE can never lie below $\mathrm{Var}(\epsilon)$, the irreducible error from (2.3).

What do we mean by the *variance* and *bias* of statistical learning method?

Varince refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set; different training data sets will result in a different $\hat{f}$. But ideally, $\hat{f}$ should not vary too much between training sets. If a method has high variance small changes in the training data can result in large changes in $\hat{f}$.

Flexible methods have hiher variance, because they fit better to the data points and changing any of the data points may cause the estiamte $\hat{f}$ to chance considerably. But for example, least squares method is relatively inflexible and has low variance, because mooving any single observation will cause only a small shift in the position of the line.(2.9)

bias refers to the error that is due to functional form of $\hat{f}$. In real life, linear relationships are very rare. So performing linear regression will result in some bias in the estimate of f. If your f is non-linear performing linear regression on different data sets will not produce an accurate estimate; so linear regression will result in high bias.

For example in 2.9 true f is non linear; so linear regression have high bias, low variance.

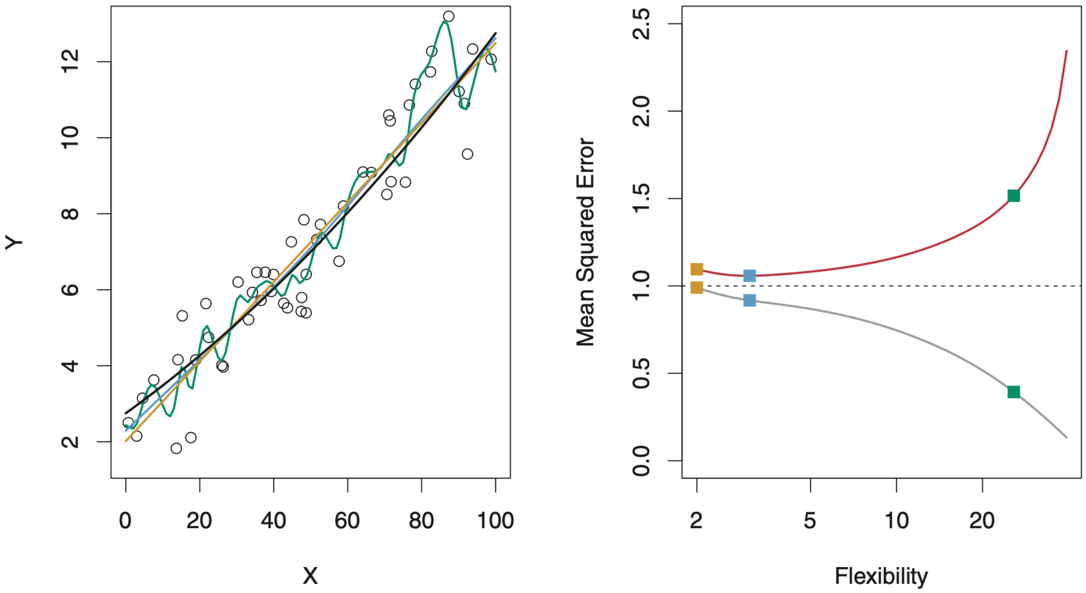

FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

In 2.10 true f is very close to linear, so linaer regression have low bias, low variance.

Generally more flexible methods result in less bias.

As a general rule, as we use more flexible methods the variance will increase and bias will decrease. The relative rate of change of these two quantities determines whether test MSE increases or decreases. As we increase the flexibility, the bias tends to initially decrease faster than the variance increases => test MSE declines. However, at some point increasing flexibility has little impact on thebias but starts to significantly increase the variance => test MSE increases.

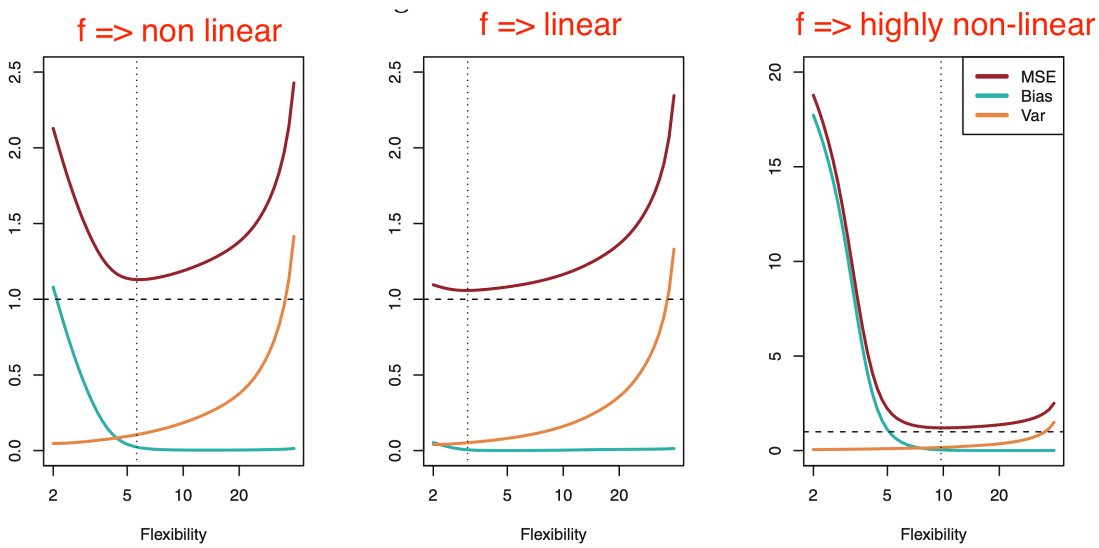

FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

Figure 2.12 shows bias and variance effect to the test MSE for different fs. Horizontal dashed line represents Var(ϵ), the irreducible error; the red curve test MSE is the sum of squared bias, variance, and variance of irreducible error. In all cases bias decreases as flexibility increaes. However, the optimal flexibility is different for each f. In the left panel the bias initially decreaes rapidly, decreasing test MSE. In center panel true f is closer to linear so there is only a small decrease in bias as flexibility increaes, and the test MSE only declines slightly before increasing rapidly as the variance increaes. Right hand panel, as flexibility increaess bias dramatically decreases because true f is very non-linear. There is alsso very little increase in variance as flexibility increases => tets MSE decreases before increasing.

This is called bias-variance trade off. Good test set performance of a method requires low variance as well as low squared bias. This is a trade off because it is easy to obtain a method with extremely low bias but high variance(for instance, drawing a curve that passes through every single training observation, or a method with low variance but high bias(by fitting a horizontal line to the data). Challange is finding a method wihch both the variance and squareed bias are low.

In real life, it is not possible to explicitly compute the test MSE, bias, or variance for methods. But we should keep this in mind.

1.3.3 The Classification Setting

So far we focused on regression setting. Problems such as bias-variance trade of also occurs in classification but in a modificated way because y_i is no longer numerical.

Suppose that we seek to estiamte f on the basis of training observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$ where now $y_1, \dots, y_n$ are qualitative.

We need to quantify the accuracy of our estimate $\hat{f}$. We can use the training *error rate*, the proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the training observations:

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

(2.8)

$\hat{y}_i$ is the predicted class label for the ith observation using $\hat{f}$. $I(y_i \neq \hat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$. If $I(y_i \neq \hat{y}_i) = 0$ then ith observation was classified correctly, otherwise it was misclassified. So (2.8) computes the fraction of incorrect classifications.

(2.8) is *training error.* But as the regression setting we are more interested in *test error* rate. The *test error* rate associated with a set of test observations of the form

$$\text{error rate}_{test} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} (I(y_{test_i} \neq \hat{y}_{test_i}))$$

(2.9)

A *good* classifier is one for which the test error is smallest.

The Bayes Classifier

We can minimize test error rate by a very simple classifier that *assigns each observation to the most likely class, given its predictor values.* In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which

$$Pr(Y = j | X = x_0)$$

(2.10)

is largest. This is *conditional probability:* it is the probabilty that $Y = j$ given the observed predictor vector x_0. This classifier is called *Bayes classifier.*

In a two-class problem where there are only two possible response values, *class 1* or *class2*, the Bayes classifier corresponds to predicting class one if $Pr(Y = 1|X = x_0) > 0.5$, and class two otherwise.

Imagine having $X = (X_1, X_2)$. For each value of X_1 and X_2 there will be a different probability of the response being class 1 or 2. For $Pr(Y = class1|X = (X_2, X_2)) > 0.5$ and $Pr(Y = class2|X = (X_1, X_2))$.

The Bayes classifier produces the lowest possible test error rate, called the *Bayes error rate*. Since the Bayes classifer will always choose the class for which $Pr(Y = j|X = x_0)$ is largest, the error rate at $X = x_0$ will be $1 - \max_j Pr(Y = j|X = x_0)$. In general, the overall Bayes error rate is given by

$$\text{Bayes error rate} = 1 - E(\max_j Pr(Y = j|X))$$

(2.11)

where the expectation averages the probabilty over all possible values of X. The Bayes eror rate is analogous to the irreducible eror.

K-Nearest Neighbors

In theory we always want to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classfier is impossible. So Bayes classifier is like a gold standard to compare other methods.

Many approaches attemp to estiamte the conditional distribution of Y given X, and then classify a given observation to the class with highest *estimated* probability. One of them is *K-nearest neighbors*(KNN) classfier.

Given a positive integer K and a test observation x_0, the KNN clasffier first identifies the K points in the training data that are closest to x_0, represented by N_0. It then esitmates the conditional probability for class j as the fraction points in N_0 whose response values equal to j.

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

(2.12)

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

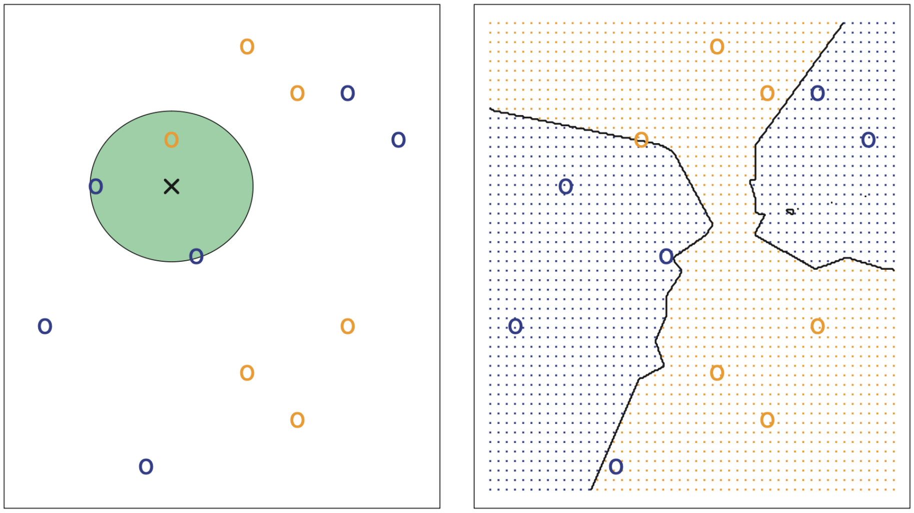

FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations.* Left: *a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue.* Right: *The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.*

Figure 2.14 provides an illustrative example of the KNN approach. Left panel => Our goal is to make prediction for the black cross point. When $K = 3$ KNN will identify the 3 oservations that are closest to the cross. There are two blue and one orange points; $Pr(Y = orange|X = x_{cross}) = 1/3$, and $Pr(Y = blue|X = x_{cross}) = 2/3$ => KNN will predict that the black cross belongs to the blue class.

The choice of K is very important. as K increases flexibility decreases => high bias, but low variance.

Just like in regression setting there is not a strong relationship between the training error rate and the test error rate.

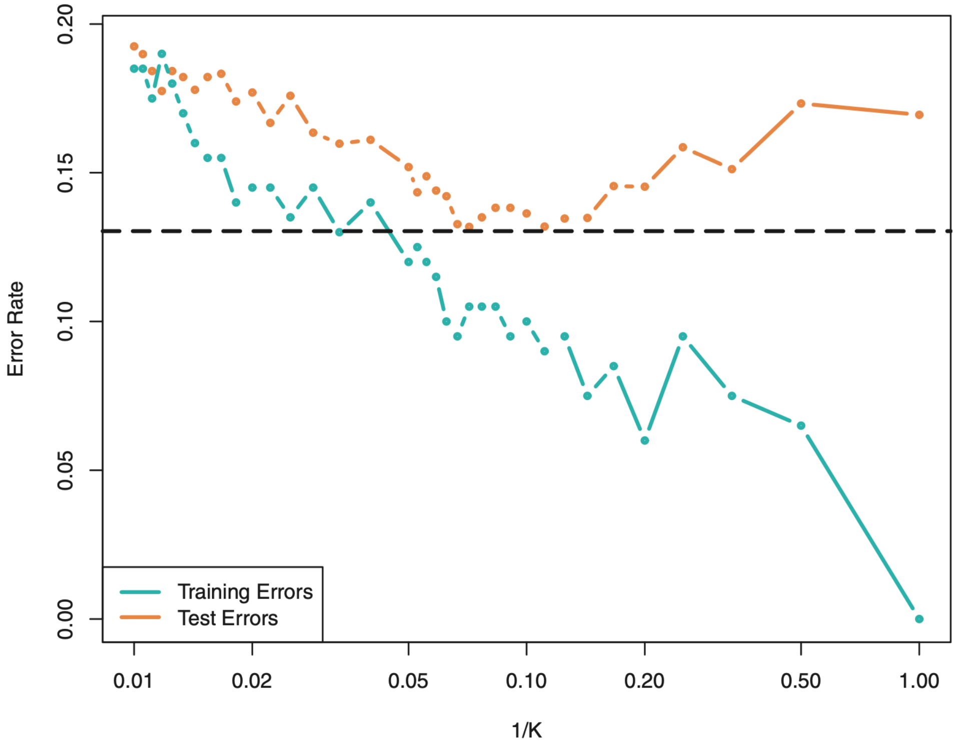

FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

as in the regression setting, the training error rate consistently declines as the flexibility$(1/K)$ increases. However, the test error rate again have a characteristic U-shape.

In both regression and classification settings, choosing the correct level of flexibility is critical. The bias-variance tradeoff => U-shape in the test error, can make this a difficult task.

2 Linear Regression

We will predict quantitative response.

```
suppressPackageStartupMessages({
library(ISLR)
library(tidyverse)
library(ggthemes)
library(sjPlot)
library(corrplot)
library(tidymodels)
library(magrittr)
library(dotwhisker)
library(hrbrthemes)
library(patchwork)
})
```

```
advertising = read_csv("./data/Advertising.csv") %>% as_tibble %>% select(-1)
advertising
```

```
# A tibble: 200 x 4
      TV radio newspaper sales
   <dbl> <dbl>     <dbl> <dbl>
 1 230.   37.8      69.2  22.1
 2  44.5  39.3      45.1  10.4
 3  17.2  45.9      69.3   9.3
 4 152.   41.3      58.5  18.5
 5 181.   10.8      58.4  12.9
 6   8.7  48.9      75     7.2
 7  57.5  32.8      23.5  11.8
 8 120.   19.6      11.6  13.2
 9   8.6   2.1       1     4.8
10 200.    2.6      21.2  10.6
# i 190 more rows
```

We are asked to suggest a marketing plan for next year which will yeild high product sales. We may want to inquire the following questions:

1. *Is there a relationship between advertising budget and sales?*

   First we should determine if there is an association between adveritisng expenditure and sales. If not, no money shpuld be spent on advertising.

2. *How strong is the relationship between advertising budget and sales?*

   If there is a relationship between advertising and sales, what is the strength of this relationship? Given a certain advertising budget, can we predict sales with a high level of accuracy? => strong relationsihp.

3. *Which media contribute to sales?*

   Do all variables–tv,radio,newspaper– contribute to sales, or just one or the two?

4. *How accurately can we estimate the effect of each medium on sales?*

   For every ollar spent on advertising in a particular medium, by what amount will sales increase? How accuretly can we predict this amount of increase?

5. *How accurately can we predict future sales?*

   For any given level of media advertisig, what is our prediction for sales, and what is the accuracy of this prediciton?

6. *Is the relationship linear?*

   If so linear regresion is appropriate tool, if not we may need to transform the predictor or the repsonse so that liner regression can be used.

7. *Is there synergy among the advertising media?*

   Does the effect of a medium on sales depend on other medium levels? Does dividing advertisement budget to two or three medium yeild a higher sales?

We can answer each of these questions using Linear regression.

2.1 Simple Linear Regression

Predicting a quantitative response Y on the basis of a single predictor variable X.

Our assumption is that there is approximately a linear relationship between X and Y; we can write this linear relationship as

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

$$Y \approx \beta_0 + \beta_1 X$$

(3.1)

For example lets say X is `TV`, and Y is `sales`

$$sales = \beta_0 + \beta_1 \times TV + \epsilon$$

or

$$sales = \beta_0 + \beta_1 \times TV$$

On (3.1) β_0 and β_1 are unknown constants that represent the *intercept* and *slope* in the linear model. Together they are known as *coefficients* or *parameters.*

We are going to use or training data to produce estimates for β_0 => $\hat{\beta}_0$ and β_1 => $\hat{\beta}_1$. Using these predicted coefficients we can predict sales;

$$\hat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV$$

or as in general form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(3.2)

2.1.1 Estimating the coefficients

Since, β_0 and β_1 are unknown, before we can use (3.1) to make predictions we must use data to estimate the coefficients. We have n observations :

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

We want our estimated coefficients to give such predictions that will fit the avaible data as well => $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$. This coefficients will allow us to draw a regression line and we want this regression line to be close as possible to the n data points we have.

There are different ways to measure *closeness.* The most common approach is minimizing the *least squares* criterion. Alternative approaches will be considered in Chapter 6.

Our predictions come from $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Then for each data we have a *residual*: difference between y and $\hat{y}$: