Class Notes

Statistical Computing & Machine Learning

Class 3

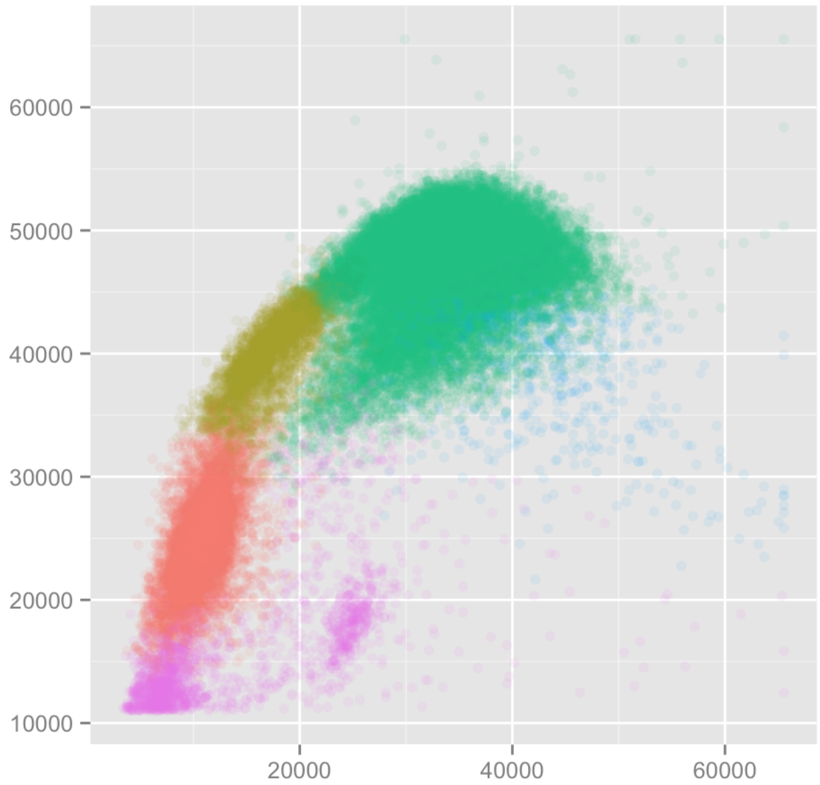
Review

- $f(\mathbf{X})$ versus $\hat{f}(\text{X})$: Platonic idea versus what we get out of the training data. Quip: "The hat means there's a person underneath the model."
- Mean Square Error — like the standard deviation of residuals
- Training vs testing data
- Smoothness, a.k.a. flexibility, model degrees of freedom
  - More flexibility → better training MSE
- Components of MSE
  1. Irreducible random noise: ϵ
  2. Bias: $f(\mathbf{X}) - \hat{f}(\mathbf{X})$
     - Caused by too much smoothness
     - Caused by omitting a relevant variable
     - Caused by including an irrelevant variable
  3. $Var(\hat{f}(\mathbf{X}))$ — how much $\hat{f}$ varies from one possible training set to another.
     - Increased by too many degrees of freedom: *overfitting*
     - Increased by collinearity and multi-collinearity.
     - Increased by large ϵ
     - Decreased by large n

Classifier

A classification setting: Blood cell counts.

Build a machine which takes a small blood sample and examines and classifies individual white blood cells.



The classification is to be based on two measured inputs, shown on the x- and y-axes.

Training data has been developed where the cell was classified "by hand." In medicine, this is sometimes called the *gold standard*. The gold standard is sometimes not very accurate. Here, each cell is one dot. The color is the type of the cell: granulocytes, lymphocytes, monocytes, . . .

- How would you go about building a classifier which uses just the x- and y- inputs?

K Nearest Neighbors architecture

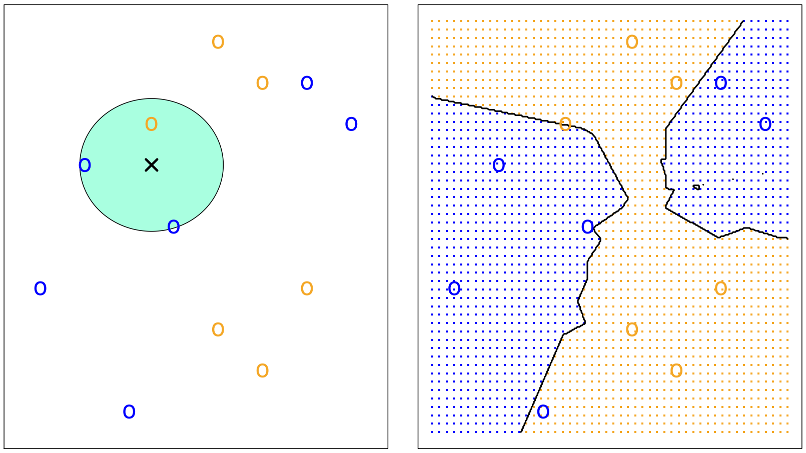


Figure 2.14 from ISL

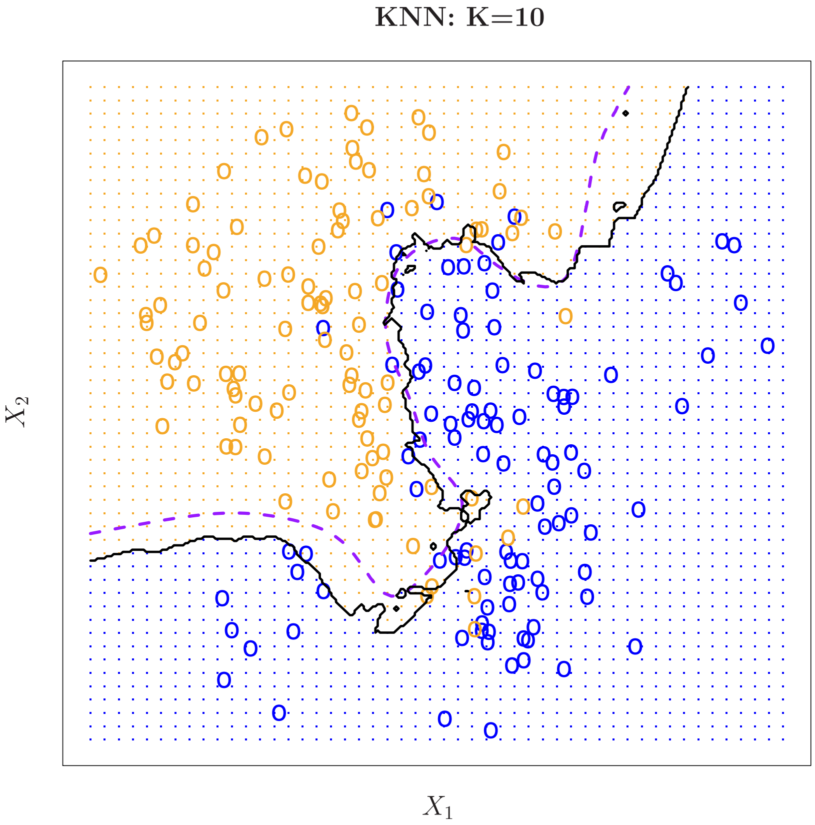


Figure 2.15 from ISL

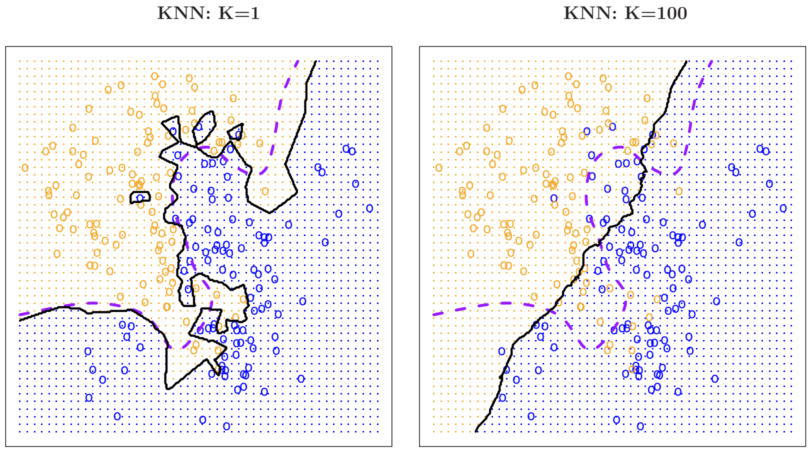


Figure 2.16 from ISL

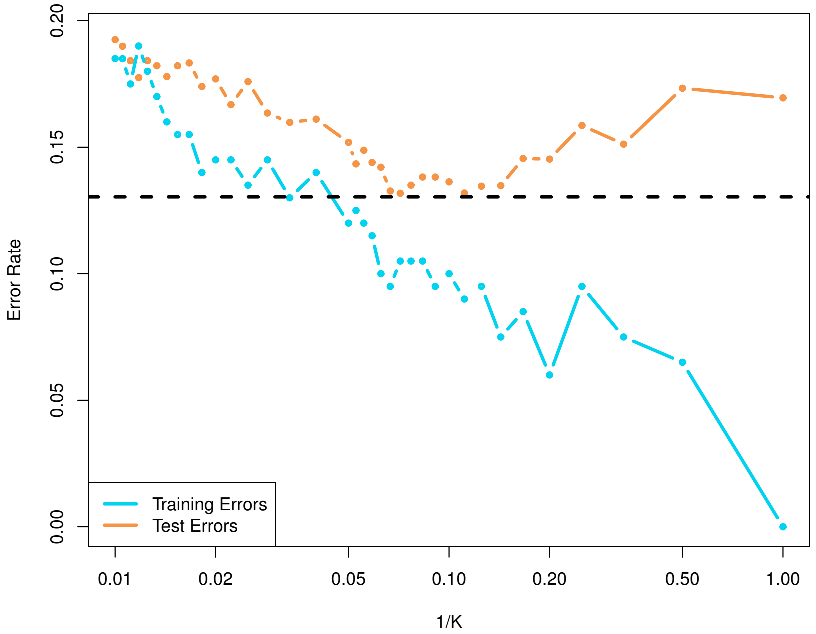


Figure 2.17 from ISL

Chapter 3

The advertising data

Basic questions:

- Is advertising worthwhile?
- If so, how to optimize its effect in sales/advertising-dollar?
- What is the overall effect? (So that the business can compare it to other ways of spending its money.)

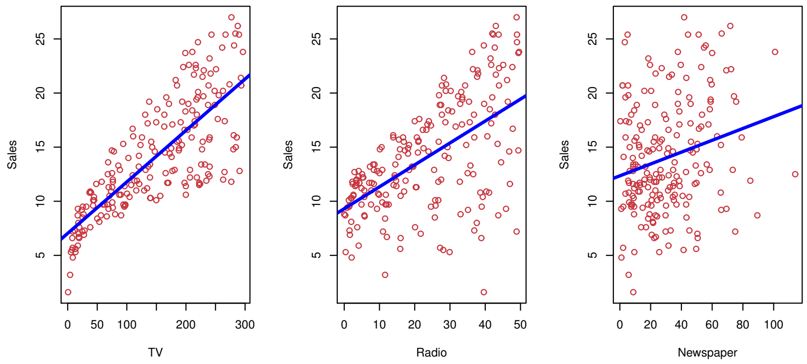


Figure 2.1 from ISL

The book poses a series of seven questions that relate more specifically to statistical techniques.

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media.

```
download.file("http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv",
    dest = "Advertising.csv")
```

```
Advert <- read.csv("Advertising.csv")
head(Advert)
```

```
##   X    TV Radio Newspaper Sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

Question 1

```
Advert$budget <- with(Advert, TV + Radio + Newspaper)
```

Technique

```
summary(lm(Sales ~ budget, data = Advert))
```

```
##
## Call:
## lm(formula = Sales ~ budget, data = Advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0546 -1.3071  0.1173  1.5961  7.1895
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept) 4.243028   0.438525   9.676
## budget      0.048688   0.001982  24.564
##             Pr(>|t|)
## (Intercept)   <2e-16 ***
## budget        <2e-16 ***
## ---
## Signif. codes:
##   0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1  ’ ’ 1
##
## Residual standard error: 2.6 on 198 degrees of freedom
## Multiple R-squared:  0.7529, Adjusted R-squared:  0.7517
## F-statistic: 603.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

What’s the answer?

Question 2

Question 3

Question 4

Question 5

Question 6

Question 7

The geometry of fitting

- Data tables: cases and variables.
- A quantitative variable is a vector.
- A categorical variable can be encoded as a set of “dummy” vectors.
- Response variable and explanatory variable
- The linear projection problem: find the point spanned by the explanatory variables that’s closest to the response. That linear combination is the best-fitting model.

  - One explanatory and the response
  - Two explanatory on board and the response on the board (perfect, but meaningless fit)
  - Two explanatory in three-space and the response (residual likely)

Measuring Accuracy of the Model

- R^2 - Var(fitted)/Var(response)
- Adjusted R^2 - takes into account estimate of average increase in R^2 per junk degree of freedom
- Residual Standard Error - Sqrt of Average square error per residual degree of freedom. The sqrt of the mean square for residuals in ANOVA

Bias of the model

- Perhaps effect of TV goes as sqrt(money) as media get saturated?
- Perhaps there is a synergy that wasn't included in the model?

Precision of the coefficients

$$\text{standard error of B coef.} = |\text{residuals}| \frac{1}{|\text{B}|} \, \frac{1}{\sin(\theta)} \, \frac{1}{\sqrt{n}} \sqrt{\frac{n}{n-m}}$$

- m — degrees of freedom in model
- θ — angle between this model vector and the space spanned by the others
- B — this model vector
- residuals — the residual vector

In-class programming activity

Day 3 activity