Ch2

June 5, 2022

1 Chapter 2

1.1 q1

a) Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.

b) Worse. A flexible method would overfit the small number of observations.

c) Better. With more degrees of freedom, a flexible method would fit better than an inflexible one.

d) Worse. A flexible method would fit to the noise in the error terms and increase variance.

1.2 q2

a) Regression and inference with $n = 500$ and $p = 3$

b) Classification and prediction with $n = 20$ and $p = 13$

c) Regression and prediction with $n = 52$ and $p = 3$

1.3 q4

a)

1. Classification 1 - Is this TV series/movie/ad campaign going to be successful or not (Response : Success/Failure, Predictors : Money spent, Talent, Running Time, Producer, TV Channel, Air time slot, etc., Goal : Prediction).
2. Classification 2 - Should this applicant be admitted into Harvard University or not (Response: Admit/Not admit, Predictors : SAT Scores, GPA, Socio Economic Strata, Income of parents, Essay effectiveness, Potential, etc., Goal : Prediction).
3. Classification 3 - Salk Polio vaccine trials - Successful/Not Successful (Response : Did the child get polio or not, Predictors : Age, Geography, General health condition, Control/Test group, etc., Goal : Prediction).

b)

1. Regression 1 - GDP Growth in European economies (Response : What is the GDP of countries predicted to be by 2050 , Predictors : Population, Per capita income, Education, Average life expectancy, Tax Revenue, Government Spending etc., Goal : Inference).
2. Regression 2 - What is the average house sale price in XXX neighborhood over the next 5 years (Response : Average house in XXX neighborhood will sell for Y next year, Z the year

after, T after that, etc., Predictors : Proximity to transit, Parks, Schools, Average size of family, Average Income of Family, Crime Rate, Price Flux in surrounding neighborhoods etc., Goal: Inference).

3. Regression 3 - Gas mileage that a new car design will result in (Response : With certain parameters being set, X is the mileage we will get out of this car, Predictors: Fuel type, Number of Cylinders, Engine Version, etc., Goal : Inference).

c)

1. Cluster 1 - Division of countries into Developed, Developing and Third World (Response : By 2050, countries in Asia can be split into these following clusters, Predictors : Per Capita Income, Purchasing power parity, Average birth rate, Average number of years of education received, Average Death Rate, Population etc., Goal: Prediction).
2. Cluster 2 - Division of average working population into income segments for taxion purposes (Response : This worker falls under this taxation bracket, Predictors : Income, Job Industry, Job Segment, Size of Company, etc., Goal : Prediction).
3. Cluster 3 - Cluster new movies being produced into ratings $G/PG/R/PG-13$ etc. (Response : This movie is a R/PG/PG-13, Predictors : Violent content, Sexual language, theme, etc., Goal : Prediction).

1.4 q6

A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f. A non-parametric approach does not assume a patricular form of f and so requires a very large sample to accurately estimate f. The advantages of a parametric approach to regression or classification are the simplifying of modeling f to a few parameters and not as many observations are required compared to a non-parametric approach. The disadvantages of a parametric approach to regression or classification are a potentially inaccurate estimate f if the form of f assumed is wrong or to overfit the observations if more flexible models are used.

1.5 q7

a)

Obs.	X_1	X_2	X_3	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)

b)

If $K = 1$ then $x_5 \in \mathcal{N}_0$ and we have

$$P\left(Y = \text{Red} \mid X = x_0\right) = \frac{1}{1} \sum_{i \in \mathcal{N}_0} I\left(y_i = \text{Red}\right) = I\left(y_5 = \text{Red}\right) = 0$$

and

$$P(Y = \text{ Green } \mid X = x_0) = \frac{1}{1} \sum_{i \in \mathcal{N}_0} I(y_i = \text{ Green }) = I(y_5 = \text{ Green }) = 1.$$

Our prediction is then Green.

c)

If $K = 3$ then $x_2, x_5, x_6 \in \mathcal{N}_0$ and we have

$$P(Y = \text{Red} \mid X = x_0) = \frac{1}{3} \sum_{i \in \mathcal{N}_0} I(y_i = \text{Red}) = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$$

and

$$P(Y = \text{ Green } \mid X = x_0) = \frac{1}{3} \sum_{i \in \mathcal{N}_0} I(y_i = \text{ Green }) = \frac{1}{3}(0 + 1 + 0) = \frac{1}{3}$$

Our prediction is then Red.

d)

As K becomes larger, the boundary becomes inflexible (linear). So in this case we would expect the best value for K to be small.