# Chapter 2

Ali Hamdani

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.

   If the sample size is large and the number of predictors are low then we would expect the flexible model to work better. This is because with more samples our $Var(\hat{f}(X_o)$ will go down. Implying less variance and error

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

   If the number of predictors is large and samples are low then the flexible statistical model runs the risk of finding the wrong relation between the variables. The variance of the function will be very high resulting in a higher test mean squared error

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

   If the relationship between the predictors and response is non-linear and unknown then a flexible statistical learning method will be better. However if its non-linear and known then non-flexible methods would work better.

   (d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.
   if the variance of the error term is extremely high then it is better to use a non-flexible model. This is because if the flexible model was chosen then it would run the risk of over fitting to the noise of the data set.

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 regression problem since salary is a continuous variable.

The goal of the application is inference since we are more interested in predictor effects than accuracy. $n = 500$ and $p = 3$

(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.

This is a classification problem since the target variable is a binomial variable (either 0 or 1). The goal of the application is prediction since we are concerned with the outcome of the product. $n = 20$ and $p = 13$.

(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. This is a regression problem since % change is a continuous variable. The goal of the application is prediction since that was written in the first line of the problem. $n = 52$ and $p = 3$

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.

This is already in the book.

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

i. **Bias**: The squared bias term is monotonically decreasing, bias is the difference of the *averages* of the estimate with its true value. In other words $(E[\hat{f}(X_o)] - f(X_o))^2$. With a flexible approach the $E[\hat{f}(X_o)]$ is going to get closer to the actual value.

ii. **Variance**: As the flexibility of an approach goes up the variance will increase monotonically. This is because as flexibility increases the model $\hat{f}$ will start fitting to the noise of the individual data set instead of fitting to actual relationship. In the equation $E[\hat{f}(X_o) - E[\hat{f}(X_o)]]$, $\hat{f}$ is going to start to deviate based on the data set. Thats why the difference is going to increase.

iii. **Training Error**: As flexibility increases the training MSE decreases. This is because on observed data the variance of $\hat{f}$ does

not matter. Bias and irreducible error are only determining component. $MSE = \sigma^2 + Bias^2$. $\sigma^2$ is **fixed** and $Bias^2$ is decreasing so MSE is also decreasing.

iv. **Test Error**: As flexibility increases the Test error will decrease at first only until an optimal level then the the model will begin to over fit the training data.

v. **Bayes Error**: This remains fixed because it is the lowest possible error rate. There is no classifier than can go below it.

4. You will now think of some real-life application 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.

   i. A model that would classify whether or not a person had cancer.

$$Y = \begin{cases} 1, & \text{Cancer} \\ 0, & \text{No Cancer} \end{cases}$$

$$X = [Age, Obesity, Smoking, Sex]$$

   The goal of the application would be inference, more specifically how do these predictors relate to the target Cancer risk.

   ii. A model that predicts whether or not a purchase on a credit card is fraudulent.

$$Y = \begin{cases} 1, & \text{Fraudulent} \\ 0, & \text{Not Fraudulent} \end{cases}$$

$$X = [Location, Time, Amount, Item]$$

   The goal of this application would be prediction. A bank would care more about capturing fraudulent activities more than inference correlated effects.

   iii. A botanist wants to classify what kind of Iris flower a particular flower is.

$$Y = \begin{cases} 0, & \text{Setosa} \\ 1, & \text{Virginica} \\ 2, & \text{Versicolor} \end{cases}$$

$$X = [Sepal_l, Sepal_w, Petal_l, Petal_w]$$

   The goal of this application would be inference since in the academic domain researchers are more interested in understanding underlying distributions than predictive power.

   (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.

i. A Realtor wants to figure out how much to sell a house.
$Y = [\$Price]$
$X = [Location, NoStories, SqaureFeet, County, State, Pool, Basement]$
This is a prediction problem since a Realtor only want to know if he either over or undersold a property.

ii. A principle wants to infer causes of students grades. $Y = [Score]$
$X = [Race, Age, Grade, Sex, Subject]$
This would be an inference problem since the principle wants to know underlying causes of a students performance.

iii. A biologist wants to know the relation between blood alcohol level and coordination. $Y = [Coordination]$
$X = [BAC]$
This would be an inference problem since the structure of the fitted function $\hat{f}$ is the most interesting.

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

i. An operator wanted to know if there was any interrelated causes of helicopter crashes from a large text corpus.

ii. Wegmans wants to know what combinations of items do customers buy together.

iii. Facebook wants to learn if there are any latent social network among their users.

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 **advantage** of using a more flexible approach is you don't need to specify the underlying relation between predictors and response variable as much. This occurs when the underlying model $f$ is non-linear and you wish to reduce the bias. This model will generally have more predictive power than the linear model. The main **disadvantages** of using a flexible approach is that the inference capability of your models decreases as you become more flexible. The margin for your approach to the noise in the particular data set also increases.

6. Describe the difference 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 disadvantages?

   Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the assumption that the population data are normally distributed. Non-parametric tests are "distribution-free" and, as such, can be used for

non-Normal variables. The main reason to choose one over the other is the amount of data. If you don't have a lot of data then using a parametric approach is more reliable. However if you have data then you can afford the flexibility of not knowing the underlying distribution of the parameters.

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, $X_1 = X_2 = X_3 = 0$.
The Euclidean distance $d_2(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ where $p, q$ are points in $\mathbb{R}^n$.

$$d_2(Obs_k, Obs_{\text{test}}) = \sqrt{\sum_{i=1}^{3}(Obs_{\text{k,i}} - Obs_{\text{test,i}})^2}$$

$Obs_1$: $\sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = 3$
$Obs_2$: $\sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = 2$
$Obs_3$: $\sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = \sqrt{10}$
$Obs_4$: $\sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{5}$
$Obs_5$: $\sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{2}$
$Obs_6$: $\sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{3}$

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

$$Pr(Y = j | X = x_o) = \frac{1}{K}\sum_{i \in \mathcal{N}_o} I(y_i = j) \tag{2.12}$$

$\mathcal{N}_o$ is the set of K closest points.
When K = 1
$\mathcal{N}_o = \{Obs_5\}$
$Pr(Y = Green | X = Obs_{\text{test}}) = \frac{1}{1}(1) = 1$
When K = 1 the probability of green is 100%

(c) What is our prediction with $K = 3$? Why?
$\mathcal{N}_o$ is the set of K closest points.
When K = 3

$\mathcal{N}_o = \{Obs_2, Obs_5, Obs_6\}$

$Pr(Y = Green|X = Obs_{\text{test}}) = \frac{1}{3}(0 + 1 + 0) = \frac{1}{3}$

$Pr(Y = Red|X = Obs_{\text{test}}) = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$

When K = 3 the probability of green is $\frac{1}{3}$ and the probability of red is $\frac{2}{3}$

(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?
The best value for k would be small because it is more flexible.