# Conceptual

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.

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

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

(d)  The variance of the error terms, i.e. σ2 = Var(ε), is extremely high.

Answer:

1. With a large number of samples, a **more flexible** method can work. There is less chance to over fit the data with such a large number of data points.
2. A small number of samples with a large number of features could lead more flexible curves to perfectly fit the training data. This however, can cause overfitting. Hence, a **less flexible** approach would work better.
3. If the relationship between predictors and response is non-linear, then we require a non-linear curve to fit the model. Thus, a **more flexible** model would yield smaller error than an inflexible linear model.
4. The variance of Error Term constitutes the Irreducible error. It is independent of predictors and hence the model predicted.
5. 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.
6. 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.
7. 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.
8. 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.

Answer:

1. This is a **Regression** problem as it involves the prediction of the response variable: the CEO Salary. Number of samples is the number of firms n = 500. The features recorded to predict the response CEO Salary are profit, number of employees and industry. So number of Features p = 3 (assuming that the factors mentioned are all the factors recorded).
2. We want to determine the **Classification** of a product as one of 2 classes: Success or Failure. Number of Samples is the number of Products n = 20. Features are price charge, marketing budget, competition price and 10 other variables. Total number of features p = 13.
3. The problem requires us to predict the percentage change in the US dollar: a **Regression** Problem. Since the data was recorded weekly over the period of a year, number of samples is the number of weeks in a year n = 52. The features recorded include % change in US, British and German Markets. Number of Features p = 3. The number of samples
4. Consider a typical set of data points for which the actual model *f* is neither too linear nor non-linear, as show in in Figure 3(a).

The corresponding Plot of Flexibility Vs Error Value is shown in 3b.

As the flexibility of the hypothesis increases, we better fit the training data points. Hence the **training error gradually reduces**.

The **variance however, increases**. With lower flexibility, the hypothesis will not vary much. As flexibility increases, subtle changes in training data can change the nature of the curve. Thus, high flexibility makes the hypothesis more susceptible to change i.e. Variance increases.

The **Squared Bias decreases**. With lower flexibility, there is great chance that we are under fitting the model. Hence a high bias. With the increase in flexibility, the predicted model becomes more complicated. There is less “simplification” of the problem (hence less bias).

The **Bayes error rate** is similar to the Irreducible error. This irreducible error **remains constant** as it does not depend on the feature vector **X** used to generate the model. This line defines the minimum test error a model can have. If the test error at any point becomes lower than the irreducible error, then we may be over fitting training data.

The **Test error** is the sum of reducible error (sum of Squared Bias and Variance of the actual model) and irreducible error. With lower flexibility, we initially have high bias and low variance, hence high Test error. With increase in flexibility, the Bias decreases and variance increases but the squared decrease in Bias is more for the typical curve. Hence Test error decreases. Once Over fitting of Data Points occurs, we get a spike in variance with low bias. Test Error, once again, increases.

1. You will now think of some real-life applications for statistical learning:
2. 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.
3. 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.
4. Describe three real-life applications in which cluster analysis might be useful.

Answer:

1. Real Life Applications of Classification:
2. **Determine whether a person has cancer**

This is an example of a binary classification problem. The patients are classified as either “has cancer” or “no cancer”. The “number of Tumors” could be a feature of determining this. The number of Features p = 1. The patient is only concerned with determining whether he/she has cancer, not the relationship between the predictor variable (number of tumors) and the response variable (the classification). Hence, the goal of this cancer classification problem is a “prediction”.

1. **Gender Classification from Name**

Given a person’s name, the objective is to determine whether the person is male or female. This is a binary classification problem with 2 response variables (“male” or “female”). Some features of the name include:

* Frequency of character “A”
* Last Letter of Name
* First Letter of Name
* Length of Name

Depending on the region, the type of features chosen may be different. Consider a bank of Indian names both male and female. Female names usually end with an “A” and tend to be longer than male names. Thus they are good features to check. The list of 4 features mentioned above is not exhaustive. Since the problem involves only determining the gender, it is a problem of “prediction”.

1. **Spam Classification**

This binary classification problem determines whether a given email is either spam/not spam. This is a well known problem in the field of Natural Language Processing. Features may include:

* Presence of the word “FREE”
* Presence of the word “Congratulations”
* Length of the Email (Spam Emails could be longer)
* Salutation (An email that does not address you by name is potential spam)

The goal is to determine if a given email is either spam or not spam. There is no concern about the relationship between constituent predictor variables and the response variable (the classification as spam/not spam). Hence, this is also a problem of “prediction”.

1. **Prediction of Visual Perception**

The goal is to determine the object a person sees. Consider a subset of say 6 objects (bottle, cat, chair, table, shoe, TV). Our model should take the fMRI scan of a person viewing one of those 6 objects at a given instant of time and predict the object category viewed. The features would be the brain activation all the voxels (3D Pixels) of the brain. This could be around 100,000 features, which can be reduced with feature selection techniques and dimensionality reduction. The response variable would be one of 6 classes (bottle, cat, chair, table, shoe, TV). Since there are p = 100,000 features and we may not have many training samples, it is important to modify the training dataset by analyzing the relationship between every feature (voxel) and group of features. However, the goal is still determining the object that the person visually perceives. Hence, the nature of the problem is that of “prediction”.

1. Three Real Life Applications of Regression
2. **Prediction of House Price**

The goal is to determine the price of a house, given certain features. This Regression Problem can be solved with the following features:

* Number of Bedrooms
* Area in Square Feet
* Does the house have a Basement (categorical field)
* Cost of living index

This predictor list is not exhaustive. The price predicted is the response variable. Since we are only concerned with predicting house price over relationship with response variables, we are dealing with a “prediction” type problem.

1. **Determine Credit Score**

The FICO scale is usually used to measure credit score. The goal is to compute credit score that ranges from 300 to 850. The major factors that can impact credit score:

* Payment History: This could be represented as rating from say 0 to 100, where 0 means one has not paid any bills on time while 100 means all payments have been made. Higher this value, higher the credit score.
* Amount Owed: As a predictor variable, it can be represented by the fraction of total available credit used. Lower the amount spent means one may have enough to pay off a future debt. This positively influences credit score.
* Length of Account History: The longer a person has kept a credit card (and payed bills), their credit score is likely to increase.

The credit score is the response variable. This computation can be used by banks and money lenders to give an individual car/house/educational loans.

1. **YouTube Analytics**

The goal is to determine the factors affecting revenue for a YouTube content creator. Some These factors include features such as:

* *Demographic viewership*: Who is watching your videos? This encompasses age groups, gender distribution and location. Content creators are paid more for views from the United States than the same from African countries.
* *Subscriber count based on region*: Greater the number of subscribers to the channel, larger is your audience and ability to expand viewership. This is turn leads to more revenue.
* *Runtime of ads on every video*: Creators are paid for every ad on their videos.
* *Duration of every video*: Longer videos tend to increase watch time.

A content creator is interested in earning money, but also constantly trying to improve the channel. They analyze how say, the number of ads are affecting their revenue. If they see a noticeable (positive) impact on increasing ads, they keep the change and look for other ways to improve. The goal is not to only estimate the revenue (the response variable) , but find ways to increase it based on its relationship with a number of features – the basic definition of an “inference” problem.

1. Three applications of clustering analysis:
2. **Recommender Systems**

Say you run an online stores that sells movies. Once a user logs in, the objective is to recommend movies that you think the user may find interesting. This can be done by monitoring activities of the user such as:

* Social Media: Using Natural Language Processing, you can analyse a user’s Tweet/ Facebook Post on a Movie to see how much they liked/hated a movie.
* Observe Star Ratings: This is explicit feedback from the user
* Purchase History: Recognize the type of movies a user has watched/purchased in the past.
* Preferred Genre

Once this information is obtained, we can group users. When plotted graphically, movie watchers closer to each other visually have similar tastes. So if Mark, John and Ross fall in the same cluster and Ross and Mark loved the movie “Imitation Game”, we would recommend John to watch the same.

1. **Document categorization**

Consider the situation where there are 10,000 documents that need to be grouped such that similar topics are grouped together. Some topics include:

* Sciences
* Technology
* Wildlife
* Movies (and more.)

Unlike classification, there is no predefined set of categories. In order to categorize documents, we could define them based on properties like:

* Title
* Keywords (The type of words used can give us an idea about the document)
* Length

When 10,000 data points plotted against a set of *N* Features **X** , we notice that similar documents are close together. A clustering technique (like K means clustering) is used to group such related data points in order to form *K Clusters*. When new document is required to be categorized, it is plotted using the N features. The cluster in which the point lies is the predicted cluster the document now belongs.

1. **Customer Segmentation (Amazon, Flip Kart)**

The goal is to understand consumer interests, keep track of site interaction, and categorize them to offer better products/services. Say we own an online store similar to Amazon. There is a large amount of data we can get from the consumer without explicit feedback:

* Cursor & Screen Position: Users are interested in different aspects of a product. Some look at the specifications, while others may pay more attention to product reviews. Knowing what a user is looking for in a product makes it easier recommend better products.
* Products Visited: Users may visit different products, but have not added anything to their cart. We can find the similarities between the products and guide users in their search.
* Products added to Cart: Certain customers on their first visit add an item to the cart, but let it go at the last minute. The next time they visit, they put the same item in their cart, but decide to change their mind (again) and not make the transaction.

Based on this data collected from every consumer, we group similar customers together. This is how “People who bought products X & Y also bought product Z”, and package offers like “Buy X+Y+Z and save $5” can be used as further incentives for users to purchase products.

1. 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?

Answer: A very flexible approach usually fits a relatively complex curve to a problem. This is advantageous to solve complex problems whose data points cannot be fitted with a simple curve (like a line in the case of linear regression). For simpler problems on the other hand, flexible curves may over fit training data. The Training error becomes insignificant, but we observe larger test error. Over fitting leads to a problem of *high variance* because simple modifications to the training data greatly alters the hypothesis f. Under normal circumstances, the hypothesis should *not* vary because of different training data.

Furthermore, such flexibility is often inversely proportional to interpretability. If the relationship between individual predictors and response variable need to be interpreted (as in the case of an inference problem) then a less flexible approach should be used. However, if only the prediction of the response variable, then a more flexible approach can be used. It can potentially lead to a higher accuracy (provided no overfitting).

1. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a para- metric approach to regression or classification (as opposed to a non- parametric approach)? What are its disadvantages?

In the parametric statistical approach, we assume the structure of the hypothesis f. The problem reduces from determining the equation to estimating the coefficients of the hypothesis. A non-parametric approach on the other hand, makes no assumptions about the equation f. This gives it the advantage of achieving a larger number of custom forms for the hypothesis which can lead to better fit, hence more accurate predictions. A caveat to this use however, is the number of data points should be large. This is required because of the large number of unknown variables in a non-parametric approach.

Depending on the level of smoothness, non-parametric approach can potentially over fit the data. This can occur in parametric analysis as well when a relatively complex hypothesis form is assumed.

7.

Answer.

1. Euclidean Distance from point to is

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation |  |  |  |  | Euclidean Distance |
| 1 | 0 | 3 | 0 | Red | 3 |
| 2 | 2 | 0 | 0 | Red | 2 |
| 3 | 0 | 1 | 3 | Red |  |
| 4 | 0 | 1 | 2 | Green |  |
| 5 | -1 | 0 | 1 | Green |  |
| 6 | 1 | 1 | 1 | Red |  |

1. With Substituting K=1 in the following equation

The test point would have the same color as the closest point near it. From the Table in 7 (a), the closest point is point 5 at a distance of . Since this point is green, the test point will also be predicted **green.**

1. With K=3, the 3 closest points to the test point are considered, The closest points are:

* Point 5 (Green)
* Point 6 (Red)
* Point 2 (Red)

The probability of the ball being Green with the given data is 1/3, while the probability of it being Red is 2/3. Hence the ball is predicted **red**.

1. The Bayes classifier gives us an optimal decision boundary. However, we are not aware of the conditional probability of the response variable Y, given a vector of predictors X. KNN gives us the implementation for real world examples by estimating this conditional probability. If the goal is to get a non-linear decision boundary, the value of K should be **smaller**. However, very low K can over fit the data points.