Homework1

Name: Chenqing Ji
Student ID: 11911303
Statistical Learning for Data Science

Due time: March 6, 2023 (Monday) 12:00am

1 Problem 1

Provide an illustration from reality for each of the subsequent concepts, along with your individual perspective.

(a) Statistical Learning Problem

(b) Data (both quantitative and qualitative)

(c) Statistic

(d) Supervised Learning

(e) Unsupervised Learning

Solution:

(a) A statistical learning problem is when we have a dataset with certain inputs and corresponding outputs, and we want to build a model that can predict the output for new inputs. An example of this is predicting housing prices based on various features of a house, such as the number of bedrooms, bathrooms, square footage, etc. My perspective on statistical learning problems is that they are a powerful tool for making predictions and uncovering insights in large datasets.

(b) Data refers to any information that can be collected and analyzed. Quantitative data is numerical data that can be measured, such as height, weight, or temperature. Qualitative data is non-numerical data that cannot be measured, such as opinions, attitudes, or beliefs. An example of quantitative data is the number of cars sold in a month, and an example of qualitative data is customer feedback on a product. My perspective is that it is essential for decision-making and can provide valuable insights when properly analyzed.

(c) A statistic is a function of a sample where the function itself is independent of the sample's distribution. The term statistic is used both for the function and for the value of the function on a given sample. Moreover, a statistic is a numerical measure that describes a sample of data, such as the mean, median, or standard deviation. Statistics can be used to make inferences about a larger population based on a sample. An example of a statistic is the average salary of employees in a company. My perspective on statistics is that they are a useful tool for summarizing and understanding data.

(d) Supervised learning is a type of statistics learning where a model is trained on labeled data, meaning the input data has a corresponding output label. The model learns to map the inputs to the correct outputs, and can then make predictions on new, unseen data. An example of supervised learning is predicting whether an email is spam or not based on its content. My perspective on supervised learning is that it can be very effective when there is a clear relationship between the inputs and outputs, but it may not be suitable for all types of problems.

(e) Unsupervised learning is a type of statistics learning where a model is trained on unlabeled data, meaning there are no corresponding output values. The model learns to find patterns and structure in the data on its own, without being given any explicit labels. An example of unsupervised learning is clustering similar customers together based on their purchasing behavior. My perspective on unsupervised learning is that it can be a powerful tool for discovering hidden patterns and relationships in data, but it can be challenging to interpret the results and make meaningful insights.

2 Problem 2

Provide an illustration for each of the subsequent two concepts to expound upon their differentiation and correlation, accompanied by your individual perspective:

(a) Sample versus Population

(b) Estimator versus Parameter

(c) Probability density function versus Histogram

(d) Statistical Learning versus Machine Learning

Solution:

(a) Sample and population are two concepts used in statistics to describe a group of individuals or observations. A population refers to the complete set of individuals or observations that we are interested in studying, while a sample is a subset of the population. For example, let's say we are interested in studying the height of all students in a school. The population in this case would be all students in the school. However, it may not be feasible to measure the height of every student in the school. Therefore, we may take a sample of students, say 100 students, and measure their height. The 100 students in this case would be the sample.From a statistical perspective, the goal is to use the information from the sample to make inferences about the population. Therefore, it is important to choose a representative sample that accurately reflects the characteristics of the population. Therefore, accurately understanding the difference between sample and population is important in statistical learning because it allows us to generalize our findings to a larger population.

(b) In statistical learning, an estimator is a formula or algorithm used to estimate the value of an unknown parameter based on a sample of data. A parameter is a characteristic of a population that we are interested in. For example, the population mean is a parameter that we may want to estimate using a sample of data. An estimator is a function of the sample data that produces an estimate of the population parameter. So, in this case the sample mean is an estimator for the population mean. The accuracy of an estimator depends on the sample size and the sampling method used. In a conclusion, estimators are important in statistics learning because they allow us to estimate unknown parameters based on a sample of data. However, it's important to choose an appropriate estimator for the parameter we are interested in estimating and to consider the potential sources of bias when interpreting the estimation results.

(c) A probability density function is a mathematical function that describes the likelihood of a random variable taking on a certain value. The area under the curve of a probability density function equals one, which means that the sum of all possible outcomes equals 100%. The probability density function is used to calculate the probability of a random variable falling within a certain range of values. A histogram is a graphical representation of the distribution of a set of data. It shows the frequency of each value or range of values in the data set. Unlike a probability density function, the area under a histogram is not equal to one. Instead, the height of each bar represents the frequency of values in the data set. In a conclusion, probability density function and histograms are both used to describe the distribution of a set of data. Probability density function are useful for calculating probabilities and making predictions about the data, while histograms are useful for visualizing the distribution of the data.

(d) Statistical learning and machine learning are two important subfields of artificial intelligence and data science. Statistical learning involves the use of statistical methods to analyze the data and make predictions or decisions. It focus on building statistical models to describe the relationships between variables in a dataset. It is often used in fields such as economics, biology, and social sciences, where understanding the underlying statistical relationships is important. Machine learning, which is designed to automatically improve the algorithm performance, is a more modern approach that uses algorithms to automatically learn patterns and relationships in data. Therefore, it is often used to make predictions or decisions in a wide range of applications, such as image recognition, natural language processing, and self-driving cars.

While both statistical learning and machine learning involve the analysis of datasets, they differ in their goals. Statistical learning typically focuses on understanding the relationship between variables and making inference based on that understanding, while machine learning focuses on developing algorithms that can learn from data and make predictions or decisions without being explicitly programmed. In a conclusion, Statistical learning and machine learning are both important subfields of data science. So understanding the differences between them is important because it can help us choose the appropriate methods and techniques for a given problem. Generally, statistical learning is often used when we have a good understanding of the relationships between data variables and want to make inference based on that relationships. However, machine learning is often used when we have a large amount of data that we need to make predictions but do not necessarily understand the relationships between data variables.

3 Problem 3

2.4.1 Solution:

(a) **Better**. Because if we use a flexible statistical learning method , a more flexible approach will fit the data closer and with the large sample size a better fit than an inflexible approach would be obtained.

(b) **Worse**. Because if we use a flexible statistical learning method in this case, it will overfit the small number of data observed.

(c) **Better**. Because in this case, the inputs and response have more degrees of freedom, so using a flexible statistical learning method can fit our requirement better.

(d) **Worse**. Because if we use a flexible statistical learning method, it can fit to the noise in the error terms and increase variance.

2.4.2 Solution:

(a) **Regression and Inference**. Because the quantitative output of CEO salary based on many factors on CEO firm's features. So we wanted to be able to explain the relationship between profit, headcount, industry, and the CEO salary, which is a regression inference problem. So, n is the 500 firms in US , and p is the profit, number of employees or industry.

(b) **Classification and Prediction**. Because in this task, we should predict new product's success or failure. Since we want to predict whether a new product will succeed or fail, this is clearly a binary classification problem, and predicting whether a product will succeed or fail requires evaluation with good predictive accuracy. So it's a classification prediction problem. So, n is the 20 similar products previously launched and p is the price charged, marketing budget, comp. price, ten other variables

(c) **Regression and Prediction**. Because in this task, we are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Therefore, we hope to predict the change trend of the result along with some factors. The more accurate the trend, the

better. Therefore, this is a regression prediction problem. So, n is the 52 weeks of 2012 weekly data and the p is % change in US market, % change in British market, % change in German market.

2.4.3 Solution:

(a) The five curves for subquestion (a) is shown in Figure 1 below.

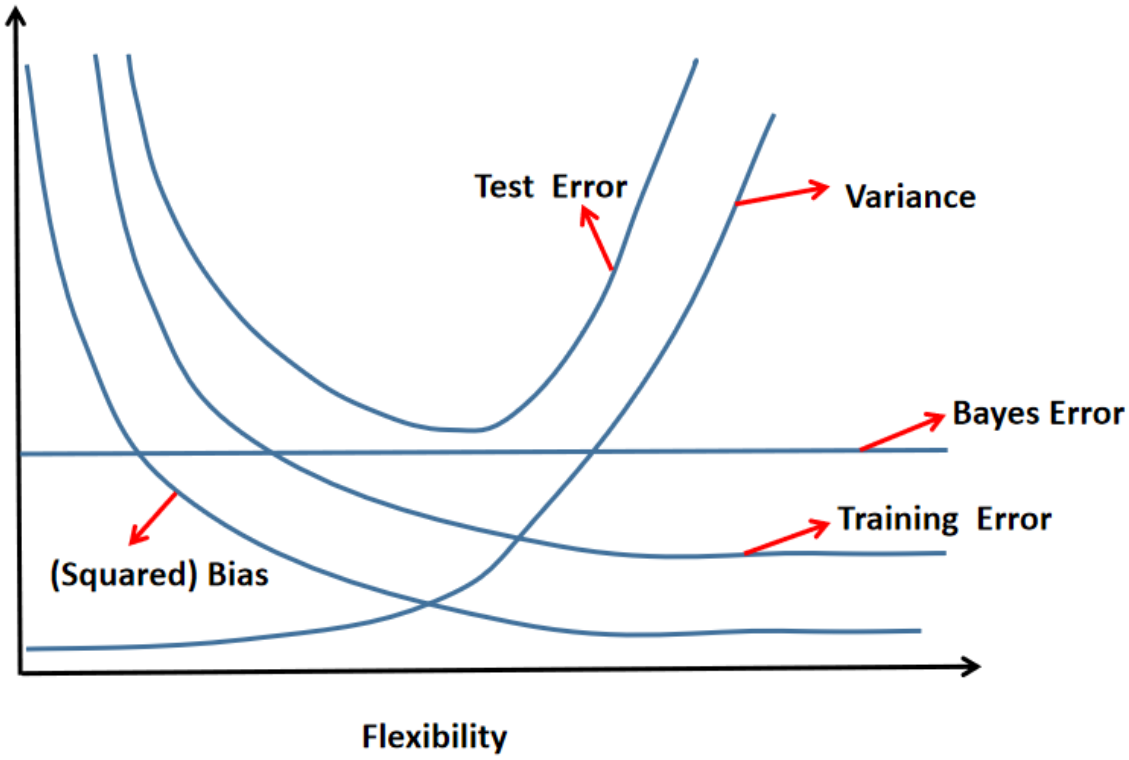

Figure 1: Solution for subquestion (a)

(b)

(1) **(Squared) Bias**: The (Squared) bias will decreases monotonically because increases in flexibility of the approaches will lead to a closer fit, which will decrease the bias monotonically.

(2) **Variance**: The variance will increase monotonically because variance represent the variability of the model's prediction for different training sets. As we move towards more flexible models, the variance will increases because the overfit.

(3) **Training Error**: As we move towards more flexible models, the training error decreases monotonically because increases in flexibility of the approaches will lead to a closer fit.

(4) **Test Error**: As we move towards more flexible models, the test error curve first decreases and reaches a minimum point. It is because that increases in flexibility of the approaches will lead to a closer fit before it overfits. After that, it increases again due to overfitting.

(5) **Bayes Error**: The Bayes error represents the irreducible error in the data. It represents the minimum error that any model can achieve, regardless of how flexible it is. When the training error is lower than the irreducible error, overfitting has taken place. As the model becomes

more or less flexible, the Bayes error remains constant, as it is determined by the inherent noise in the data.

2.4.4 Solution:

(a)

(1) **Medical Diagnosis**: Medical diagnosis is a crucial area where classification is used widely. The goal of medical diagnosis is to predict whether a patient has a specific medical condition, so it is the goal of prediction. The predictors in medical diagnosis can be clinical tests, symptoms or medical history. The response of this application is a binary outcome: either the patient has some illnesses or not.

(2) **Agricultural Planting**: In this application, we predict whether an existing crop is healthy or not based on certain characteristics it exhibits. Therefore, the response is the type of disease that affects a plant, and the predictors are features of the plant such as leaf color, texture, and shape. The goal of this task is inference– to infer the type of disease that affects a plant, based on the characteristics of the plant.

(3) **Industrial Production**: In this application, We will predict whether there is damage in a batch of new parts according to some features of the surface of the product parts such as the degree of smoothness. Therefore, the response is Whether a part is damaged or not which is a binary classification problem. The predictors are some features of the surface of the product parts such as the degree of smoothness. The goal of this task is prediction– When a new batch of parts arrives, we use this surface information to predict roughly how many broken parts are inside.

(b)

(1) **Medical Research**: Regression analysis can be used in medical research to determine the relationship between a disease and its risk factors. In this case, the response is a specific disease, and the predictors could be many factors such as age, sex, family history and genetic factors. The goal of this application is inference: to understand the relationship between a certain disease and many risk factors.

(2) **Credit Analysis**: Credit analysis is the process of determining the creditworthiness of an individual or organization. Regression analysis can be used to predict the credit score of an individual based on various factors such as income, credit history, employment status, etc. The response is the credit score of an individual and the predictors could be various factors such as income, credit history, employment status, etc. The goal of this application is prediction: to predict the credit score of an individual based on the predictors.

(3) **Financial Analysis**: Regression analysis can be used to model the relationship between a company's financial performance and various financial metrics such as revenue, expenses, and profits. In this case, the response could be a financial metric such as profitability or revenue, and the predictors could be factors such as marketing spend, cost of goods sold, and inventory levels. The goal of this application could be prediction: to predict the profitability or revenue of a company based on the predictors.

(c)

(1) **Image Segmentation**: Cluster analysis can be used to group pixels in an image based on their color, texture, and brightness. By grouping similar pixels into clusters, image segmentation can help to identify and separate objects from the background, allowing for more accurate object recognition and analysis.

(2) **Fraud Detection**: Cluster analysis can be used to identify unusual patterns or behaviors in financial transactions that could indicate fraud. By grouping similar transactions into clusters, fraud detection algorithms can identify transactions that deviate significantly from the norm and flag them for further investigation.

(3) **Market Segmentation**: Cluster analysis can be used to segment customers into distinct groups based on their preferences, behaviors, and demographics. This can help companies tailor their marketing strategies and product offerings to specific customer segments, improving customer satisfaction and retention.

2.4.5 Solution:

In statistical learning, flexibility refers to the degree to which a model can fit complex patterns in the data. Therefore, a more flexible model can fit complex patterns, while a less flexible model is more constrained in the patterns it can capture. So, The advantages for a very flexible approach for regression or classification are obtaining a better fit for non-linear models. However, the disadvantage is that it may overfit the training data, perform poorly on unseen data and follow the noise too closely.

In regression or classification, a more flexible approach might be preferred when there is a complex relationship between the input features and the output variable.For example, when we are interested in prediction and not the interpretability of the results.If the input features have a non-linear relationship with the output variable, a more flexible model such as a neural network or decision tree might be able to capture this relationship better than a less flexible model such as linear regression.

On the other hand, a less flexible approach might be preferred when the number of input features is small or when the relationship between the input features and the output variable is simple. In these cases, a less flexible model such as linear regression or logistic regression might be able to provide accurate predictions while being less prone to overfitting.

2.4.6 Solution:

Parametric and non-parametric are two different types of statistical learning approaches. In a parametric approach, the model assumes a specific functional form or shape for the underlying relationship between the independent and dependent variables. In contrast, a non-parametric approach makes no assumptions about the underlying functional form and instead estimates the relationship between the variables base on a very large number of observations.

Advantages of a parametric approach to regression or classification:

(1) Parametric models are often easier to implement and interpret compared to non-parametric models.

(2) Parametric models tend to be more efficient, requiring less data to produce accurate results compared to non-parametric models.

(3) Since parametric models have a pre-defined structure, the model parameters are usually interpretable, making it easier to draw meaningful conclusions about the relationship between the variables.

Disadvantages of a parametric approach to regression or classification:

(1) Parametric models are sensitive to the distributional assumptions made about the data. If the assumptions are incorrect, the model may produce biased or inaccurate results.

(2) Parametric models may not be flexible enough to capture complex relationships between the variables.

(3) The performance of a parametric model can be limited by the complexity of the assumed functional form. If the actual relationship between the variables is more complex than the model assumed, the model may perform poorly.

2.4.7 Solution:

(a) The solution is shown in Figure 1 below.

Obs.	X1	X2	X3	Distance(0, 0, 0)	Y
1	0	3	0	3	Red
2	2	0	0	2	Red
3	0	1	3	$\sqrt{10}$	Red
4	0	1	2	$\sqrt{5}$	Green
5	-1	0	1	$\sqrt{2}$	Green
6	1	1	1	$\sqrt{3}$	Red

Figure 2: Solution for subquestion (a)

(b) **Green**. Because the observation 5 is the closest neighbor for K = 1.

(c) **Red**. Because observations 2, 5, 6 are the closest neighbors for K = 3. Observations 2 is Red,5 is Green, and 6 is Red. Since red is predominant in its neighborhood, it is predicted to be red when K is 3.

(d) **Small**. This is because a smaller value of K leads to a more flexible decision boundary that can better capture the complex nonlinear relationships between the input features and the output labels. However, a large K will fit a more linear boundary for which it takes more points into consideration.