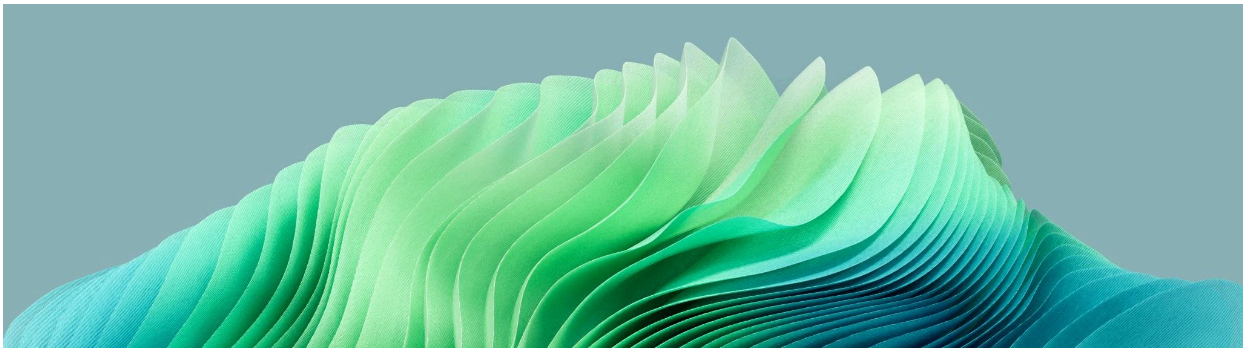


Conceptual Questions

Chapter 2 : Statistical Learning

Q1. 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.

When n>>p, that means we have ample observation and the dimensions are very low meaning we won't be having the curse of dimensionality. A highly flexible model comes with high variance but this limitation is constrained and effectively managed, as the sheer amount of data helps the model to recognize generalized patterns that will help predict future values and not stray to random noise, stopping the case of overfitting. Hence, a more flexible model is a **better** than inflexible

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

When p>>n, it generally leads to the case of overfitting in highly flexible models, as with more p the flexible method takes the form of polynomials with high dimensions like x^100,or x^300. With so many features the data is spread across very sparse or thin (curse of dimensionality), making the requirement of more data exponentially high to cover the space, as it NEEDS to be covered because highly flexible models will learn not only the generalized patterns but also all the random noise patterns making them work exceptionally well on training data but fail on validation and test data, also called as overfitting.
Methods relying on distances like K-nearest neighbors (KNN) also fail because of the thinly spread sparse data
Hence, when p>>n, a highly flexible model leads to overfitting and curse of dimensionality making them a **worse** fit than inflexible statistical learning models.

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

Highly flexible models are generally non-linear. They have more parameters and can therefore learn more complex, non-linear relationships in the data. This would make flexible models generally a **better** fit than inflexible methods.

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

MSE (Mean square error) = var(x) + Bias^2 + Var(e)
Here variance = E(x-mean)sq making this a positive value, Bias = E(fcap(x)-f(x)) hence squaring it ensures a non negative value , considering these we can say that the lowest value of MSE will always be Var(e) which is the irreducible error or noise that is out of our control.
Now according to our question, the variance var(e) is extremely high, making the lowerbound value of MSE extremely high, which we do not want as we always tend to make MSE value as low as possible. Moreover, flexible models already have high var(x) and now with high var(e) the overall MSE would be extremely high for test or validation dataset hence, Flexible statistical learning model will be **worse** than inflexible statistical learning model.

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** as CEO salary is a quantitative value, where we are more interested in **Inference**, here **n = 500 and p = 3** (profit, number of employees, industry), response variable= 1(CEO Salary)

(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.

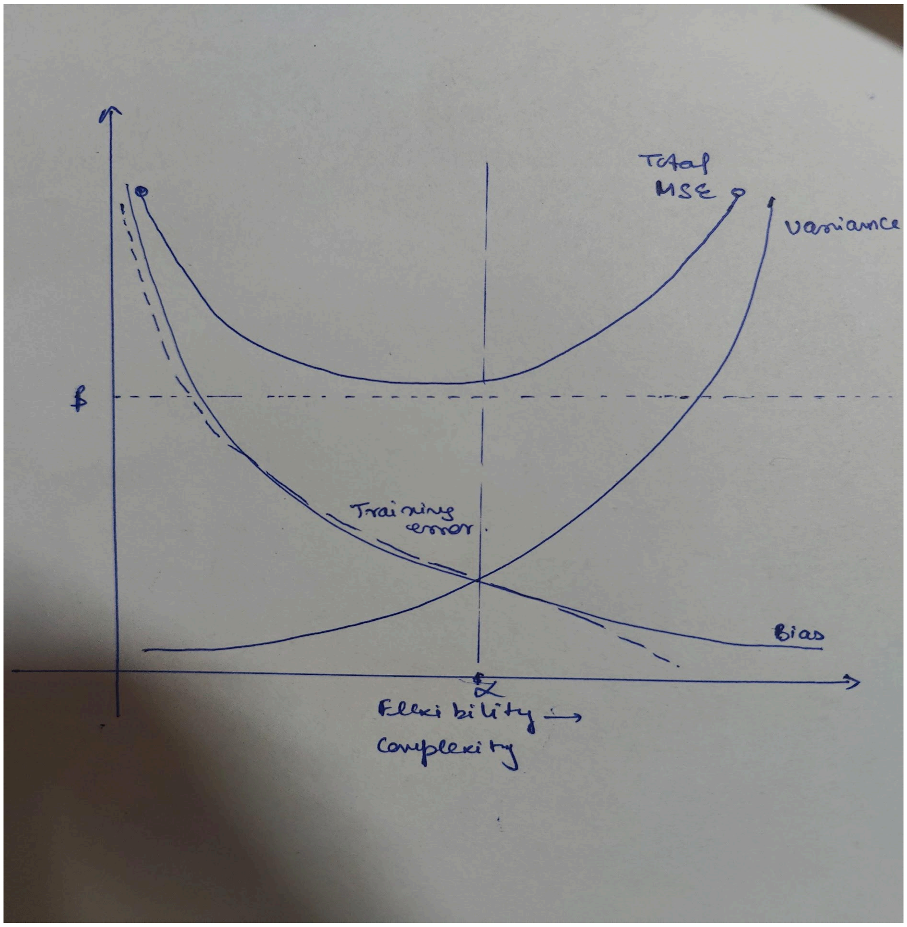
n=20, p=13, response variable =1 (Success or failure) ; this is also a **classification problem** but we are more interested in **prediction.**

(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

n=52, p=3, response variable = 1(%change in USD/Euro exchange rate); we are interested in **prediction** but this is a r**egression problem** as we are predicting a continuous value rather than a class.

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.



There are 6 parts in this graph;

1. **Bias**2 :** - the curved continuous line which is first high then lowers down
2. **Variance :** contrary to Bias, it starts of low and ends up at higher value
3. **Irreducible error Var(e)**- displayed by the dashed horizontal line
4. **Test error/ MSE** : Continuous curve whose value goes down, then goes up at the end, it does not go down below irreducible error
5. **Training error:** displayed by dashed curve line which starts of high and then as flexibility increases it goes down
6. **Optimal complexity:** Denoted by the dashed vertical line where MSE curve at its lowest point.

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

We will go each curve one by one

1. **Bias**2-** Bias is defined as the difference between the curve estimated by the chosen statistical method and the actual true function (which inherently is a theoretical ideal function), Therefore whenever the flexibility or complexity of the statistical model increases, the model tends to incorporate more values, random noises of the dataset , aka it learns the

dataset more closely, which makes it resemble true function more hence reducing the Bias (this can also lead to overfitting)

2. **Variance-** This is defined as the sensitivity of the function or curve when different datasets are used, meaning , in ideal situations we would want variance to be as low as possible , we would want all the curves obtained from different datasets so be similar so as to obtain one generalized function that would represent all the training , validation and test dataset , but if introducing new values changes the curve a lot? That means there is a lot of variance in the statistical model.
    a. **Now looking at the diagram**, when the flexibility of the model is very low, the variance is very low, simultaneously when the complexity increases, so does variance as with more flexibility it is able to memorize complex patterns of the training dataset, so much that it can learn even minor intricacies in patterns, therefore introducing slightly different data points would then yield significantly different resulting model, leading to a high variance
3. **Irreducible Error -** This is the error that we can not control, which exists in real world, when considering the variance tradeoff equation which is **MSE (Mean square error) = var(x) + Bias^2 + Var(e),** we ideally want variance (curve sensitivity) and Bias(our model- true f) to be as low as possible , and since both the values are positive , the lowest they can ideally go is 0. Making MSE = var(e) ; or var(e) being the lowest value MSE can go , now when you look at the diagram , the MSE curve never goes bellow the dashed line.
4. **Test error/MSE- The error on Test dataset,** This curve is the combination of all the variance , bias and irreducible error curve , notice how it never goes below the irreducible error? Thats explained in the above para
5. **Training Error -** This is the average error on the training data, which starts of high on low complexity but when the model becomes more flexible , it learn the training data , as well as the noise more effectively, reducing the error and the bias to an inherent 0
6. **Optimal complexity -** This is the point where MSE curve is at the lowest

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

Customer loan default risk at a finance consulting company- A finance consulting company can develop a classification model for a bank to assess the credit risk of loan applicants. By predicting whether an applicant is likely to default on a loan, the bank can make informed decisions.

- **Response**: The response variable is the binary outcome of whether a customer defaults on a loan e.g "Default" or "Not Default", “yes” or “no” etc.
- **Predictors**: Relevant predictors include the applicant's credit score, annual income, debt-to-income ratio, length of credit history, employment status etc

- **Primary goal is prediction**. The bank is most concerned with accurately forecasting which new applicants will default to minimize financial risk and losses.
- **Secondary goal** can be inference, While understanding which factors correlate with default risk is also useful, the immediate business objective is the predictive accuracy of the model.

Classifying mental health risk- In clinical psychology, a classification model can be used to help identify individuals at a higher risk of developing a specific mental health condition, such as depression or anxiety, based on their self-reported symptoms and historical data.

- **Response**: The response variable is a categorical classification of mental health risk e.g "High Risk," "Moderate Risk," "Low Risk"
- **Predictors** can include self-reported patient symptoms like frequency of sadness, sleep patterns, energy levels, demographic information (age, gender), family history of mental illness, and past psychological assessments etc
- **The main goal is inference**. While predicting a person's risk level can be a critical outcome, the psychologist's primary objective is often to understand the underlying relationships between the predictors and the mental health risk so as to correctly diagnose the patient. A model that explains the factors on why the patient is at high risk is more beneficial as to a model that can only tell whether the patient will be at a high risk or not for then the doctor can use the inference data for a more personalized treatments for different patients depending on models result.

Overwatch Character type detection- A developer or analyst working with Overwatch might use a classification model for a computer vision project to identify which character, or "hero," is visible in a game screenshot. This could be used for automated game analysis or tracking hero usage patterns.

- **Response**: The response variable is the specific hero character detected in the image, such as "Genji," "Reinhardt," or "Mercy". This is a multi-class classification problem.
- The predictors are the **pixel data of the image** itself. The model analyzes visual features within the image, such as colors, shapes, and character models, to make a classification.
- The **goal is prediction**. The model's purpose is to accurately and automatically classify the character in any new image it receives.

(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.

Optimizing advertising spend in marketing - A marketing team wants to know how much their advertising efforts translate into sales. They can use multiple regression to build a model that quantifies the impact of different advertising channels on revenue. This helps them allocate their budget more effectively to maximize return on investment (ROI).

- **Response**: The response variable is the numerical value of sales revenue.
- **Predictors**: Predictors include the amount of money spent on different marketing channels, such as TV, radio, and digital advertising, as well as potential factors like seasonality or competitor pricing.
- The **primary goal** for this application is **inference**. While a marketer wants to predict future sales, their more critical task is understanding the *relationship* between each marketing channel and sales. For example, the regression coefficients reveal which channel has the strongest positive impact and by how much. This insight allows the company to understand the underlying mechanics of their sales and make informed strategic decisions, like shifting more of their budget to the most effective channels.

Predicting competitive skill rating (SR) in OW- A data analyst could use regression to model and predict a player's competitive skill rating (SR) based on their in-game performance metrics. This could help with matchmaking accuracy or player recruitment.

- **Response**: The response variable is the player's numerical skill rating (SR).
- **Predictors**: Predictors could include a wide variety of in-game statistics, such as:
    - **Offensive metrics**: Damage per minute, final blows, and solo kills.
    - **Defensive metrics**: Damage blocked, healing done, and objective time.
    - **Hero-specific stats**: Such as "Shatter kills" for Reinhardt, "Resurrections" for Mercy. “Pulls” for Lifeweaver or Orb kills for symmetra
- The primary goal is **prediction**. The system's main objective is to accurately forecast a player's SR based on their performance metrics. An accurate prediction can be used by the matchmaking algorithm to create balanced teams or for a team manager to identify high-potential players, without needing a deep causal understanding of *why* certain metrics lead to a higher SR. The focus is on the accuracy of the final predicted value.

Analyzing drug effectiveness- A pharmaceutical company can use regression analysis to study the effectiveness of a new drug. Researchers can model how different dosages of a drug affect a patient's blood pressure.

- **Response**: The response variable is the patient's blood pressure, a continuous numerical measurement.
- **Predictors**: drug dosage, patient's age, weight, and other relevant health metrics.
- The **primary goal is inference**. The researchers are most interested in understanding the relationship between drug dosage and blood pressure. By analyzing the coefficients of the regression model, they can quantify exactly how much a change in dosage affects blood pressure and determine if this relationship is statistically significant. While predicting a specific patient's blood pressure is a possible use case, the core scientific goal is to establish and understand the cause-and-effect relationship.

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

Cluster analysis is implemented when the given data is unsupervised, meaning it has no response variable so the algorithm makes clusters based on patterns in the data. One of the most common methods is the k means clustering method

Analyzing player behavior and strategies

In *Overwatch*, game developers can use cluster analysis to identify different playstyles and strategic tendencies among players or teams. By analyzing in-game metrics, they can create player archetypes that help with matchmaking, game balancing, and competitive strategy development.

- **Variables**: Key variables for clustering could include statistics such as:
  - **In-game performance:** Damage dealt, healing provided, objective time, and total kills.
  - **Character usage:** The frequency and combination of different "heroes" played.
  - **Strategic actions:** Position on the map, timing of ultimate abilities, and engagement patterns.
- **Resulting Clusters**: The clusters might reveal distinct player archetypes, such as:
  - "Aggressive Pushers" who prioritize quick, offensive plays.
  - "Defensive Sentinels" who focus on protecting objectives and holding ground.

    - "Supportive Utility" players who prioritize healing and using abilities to enhance their team.

The clusters inform developers on how to design new heroes or adjust existing ones to ensure game balance. An esports coach could analyze the clusters of opposing teams to prepare countermeasures for their specific strategic profiles.

2. Stationery company, Customer segmentation

A stationery company can use cluster analysis to segment its customer base. By grouping customers with similar purchasing behaviors, preferences, and demographics, the company can create targeted marketing campaigns and develop products that appeal to specific segments.

- **Variables**: Useful variables for clustering customers could include:
    - **Purchasing habits**: Average order value, frequency of purchases, product categories purchased (e.g., planners, pens, art supplies).
    - **Demographics**: Age, location, and subscription status.
    - **Engagement**: Response to past promotions, website browsing behavior, and loyalty program participation.
- **Resulting Clusters**: The analysis might identify segments such as:
    - "High-Value B2B Clients" who place large, infrequent orders for office supplies.
    - "Creative Hobbyists" who frequently purchase a variety of specialized art and craft materials.
    - "Budget-Conscious Students" who buy basic pens and notebooks at a low price point.

The marketing team can design a personalized email campaign for the "Creative Hobbyists" segment featuring new art supply arrivals, or offer a bulk discount to the "B2B Clients" to incentivize larger purchases.

Identifying patient profiles

In healthcare, cluster analysis can be used to identify groups of patients with similar medical characteristics, symptom profiles, or risk factors. This can help in tailoring treatment plans, predicting disease progression, and personalizing patient care.

- **Variables**: Variables for clustering patients could include:
    - **Clinical data**: Blood pressure, cholesterol levels, lab results, and patient history.
    - **Lifestyle factors**: Diet, exercise habits, and smoking status.
    - **Symptom profiles**: Frequency and severity of reported symptoms.
- **Resulting Clusters**: The clustering might reveal distinct patient profiles, such as:
    - Patients with a high risk of cardiovascular disease based on a specific combination of blood pressure and lifestyle factors.
    - Patients with similar symptom patterns for a rare disease, which can aid in diagnosis.

Medical professionals can use these patient clusters to develop standardized, evidence-based treatment protocols for each group. For example, a doctor might prescribe a specific care regimen to the "high risk cardiovascular" cluster, improving overall patient outcomes.

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?

There is no universal model that is a better fit for a classification or a regression problem. A very flexible model like deep neural or high degree polynomial regression can fit more complex non linear data. Whereas a less flexible approach makes stronger assumptions about the relationship between response and predictors , the core tradeoff between them is the bias ,variance and the primary goal of whether we need to perform an inference or a prediction.

	More flexible model	**restrictive/ inflexible model**
Advantages	- Is better suited for **prediction** - Has **low bias**, making it a better fit	- Is better suited for **Inference** - **Computationally cheaper** because of lower parameters - Less prone to overfitting as they have lower variance and small datasets
Disadvantages	- Has high variance which makes them very sensitive to external data leading to classic case of **overfitting** - Needs a lot of training data so as to generalise true patterns - **Lower Interpretability** making it difficult to understand the underlying reasons - Can be more **computationally expensive** to train these models	- Has higher bias, which can make it too simple to capture the true underlying relationship resulting in **underfitting** - Poor predictive performance as it does struggles with complex non linear data
Examples:	- Deep neural networks, complex splines , decision trees	- Linear regression

A more flexible approach is generally preferred when:

- **High complexity and non-linearity:** When the true relationship between the predictors and the response is known to be complex and highly non-linear, a flexible model can better capture these intricate patterns.
- **Large sample size (n):** With a large number of observations, there is enough data to constrain a flexible model, reducing the risk of overfitting and high variance.
- **Prediction is the priority:** If the ultimate goal is to produce the most accurate predictions possible for new data and interpretability is a secondary concern, a flexible model is often the better choice

When to prefer a less flexible approach

- **Inference is the primary goal:** When the goal is to understand the relationship between the predictors and the response (inference), a less flexible model (e.g., linear regression) provides greater interpretability. This is crucial in fields like social science or medicine where understanding *why* a variable affects the outcome is paramount.
- **Small sample size (n):** With a limited number of observations, a flexible model is likely to overfit the noise in the training data. A less flexible model, by making simplifying assumptions, can provide more stable and reliable results on unseen data.
- **Low dimensionality and simplicity:** When the number of predictors (p) is very large but the number of observations (n) is small (the "curse of dimensionality") or when there is reason to believe the underlying relationship is simple, a less flexible model is less likely to overfit and may perform better.

6. Describe the differences 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 nonparametric approach)? What are its disadvantages?

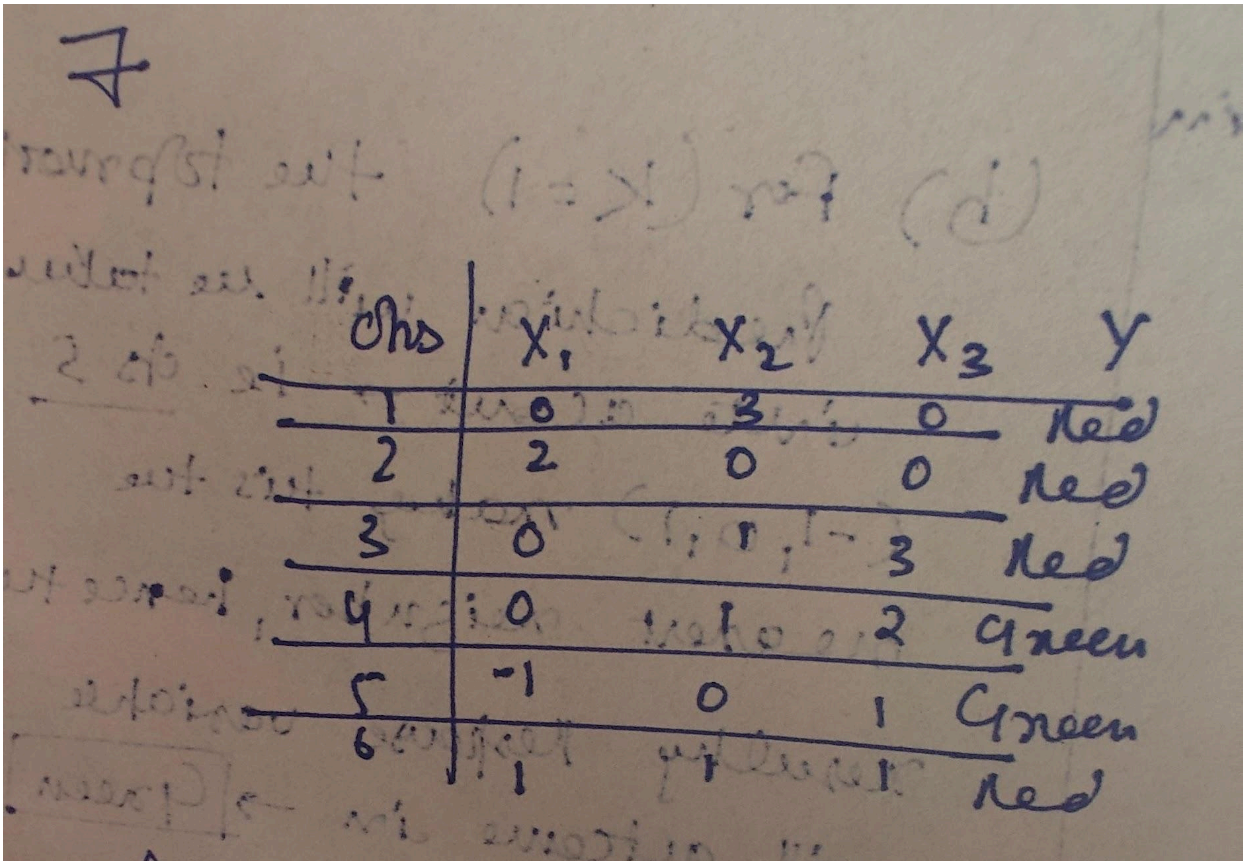
The core difference between parametric and non parametric approaches are that the former assumes the functional form of the statistical learning function whereas the latter does not.

The **Parametric approach assumes a specific predefined functional form** for the relationship between predictors and the response. The model is then defined by the fixed finite number of parameters which is estimated from the training data.
For example a **linear regression is a parametric approach** because it assumes a linear relationship Y≈β0+β1X1+...+βp. The model here is defined by the parameters aka the coefficients $\beta 0$,$\beta 1$,...,βp determined by the data.

A **non parametric approach does not make assumptions** about the functional form of the relationship, the number of parameters are not fixed and can grow with the data
For example **KNN**, simply predicts the response for a new observation based on the responses of its k nearest neighbors in the training data

The **advantages of parametric approach** are that it is computationally cheaper and easier to interpret, data efficient - can perform well with small data , However they have certain disadvantages like, HIGH Bias or the risk of misspecification where a poor approximation of the true form will lead to For instance, if a linear relationship is assumed for a highly non-linear problem, the model will fail to capture the real complexity.
Less flexible as they already assume a function and ar**e highly dependent** on them.

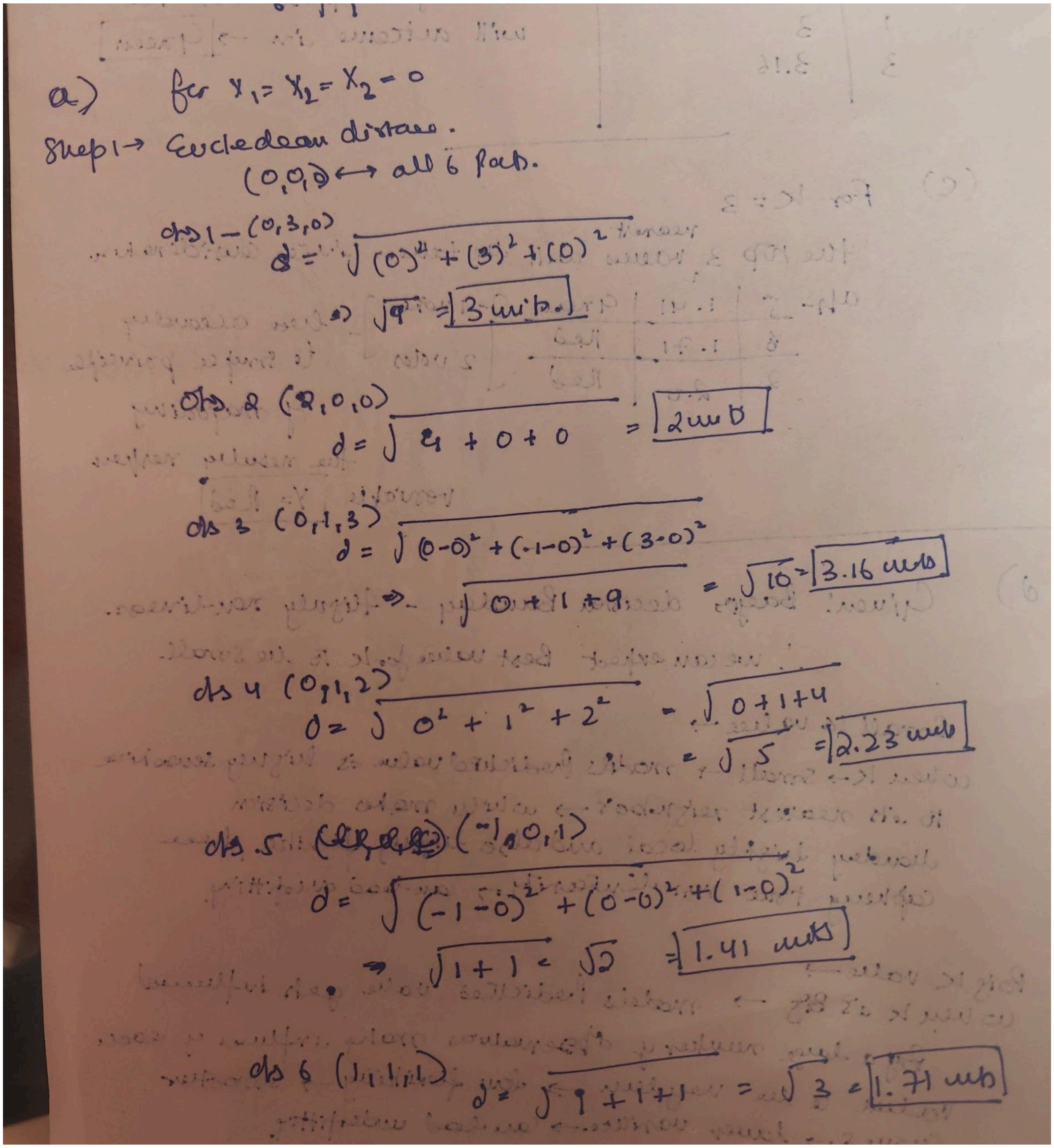
7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.



7

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

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.



a) for $X_1 = X_2 = X_2 = 0$

Step1 → Eucledean distance.

(0,0,0) ↔ all 6 Obs.

obs 1 – (0,3,0)

$$d = \sqrt{(0)^2 + (3)^2 + (0)^2}$$

$$\Rightarrow \sqrt{9} = \boxed{3 \text{ units.}}$$

obs 2 (2,0,0)

$$d = \sqrt{4 + 0 + 0} = \boxed{2 \text{ units}}$$

obs 3 (0,1,3)

$$d = \sqrt{(0-0)^2 + (-1-0)^2 + (3-0)^2}$$

$$\Rightarrow \sqrt{0 + 1 + 9} = \sqrt{10} = \boxed{3.16 \text{ units}}$$

obs 4 (0,1,2)

$$d = \sqrt{0^2 + 1^2 + 2^2} = \sqrt{0+1+4} = \sqrt{5} = \boxed{2.23 \text{ units}}$$

obs 5 (-1,0,1)

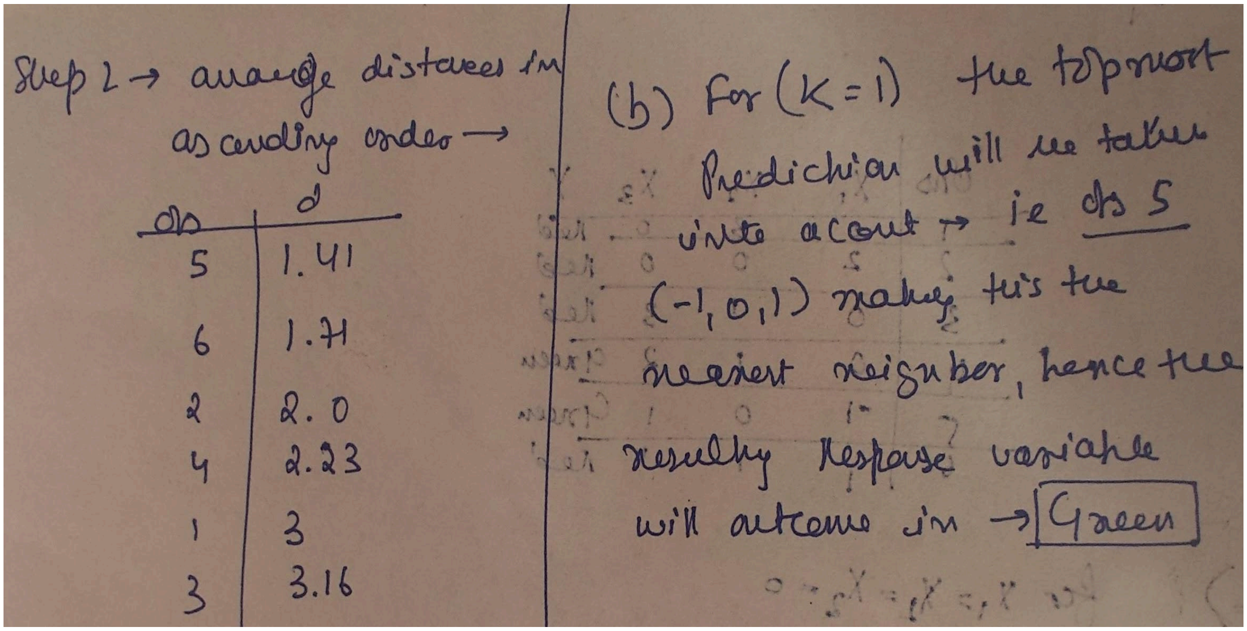
$$d = \sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2}$$

$$\Rightarrow \sqrt{1+1} = \sqrt{2} = \boxed{1.41 \text{ units}}$$

obs 6 (1,1,1)

$$d = \sqrt{1+1+1} = \sqrt{3} = \boxed{1.71 \text{ units}}$$

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

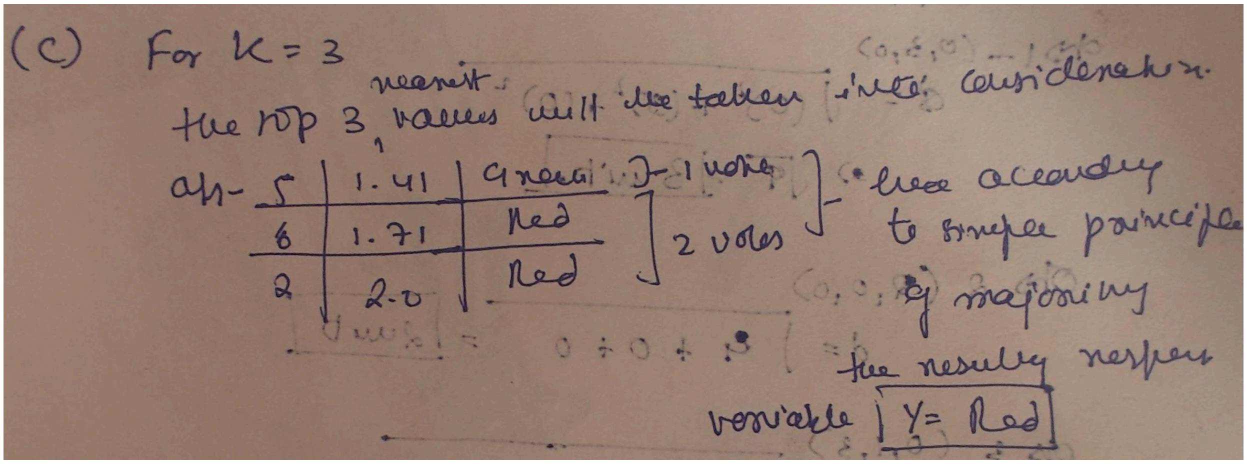


Step 2 → arrange distances in ascending order →

obs	d
5	1.41
6	1.71
2	2.0
4	2.23
1	3
3	3.16

(b) For (K = 1) the topmost Prediction will be taken into acount → ie obs 5 (-1, 0, 1) making this the nearest neighbor, hence the resulting Response variable will outcome in → $\boxed{\text{Green}}$

(c) What is our prediction with K = 3? Why?



(c) For K = 3

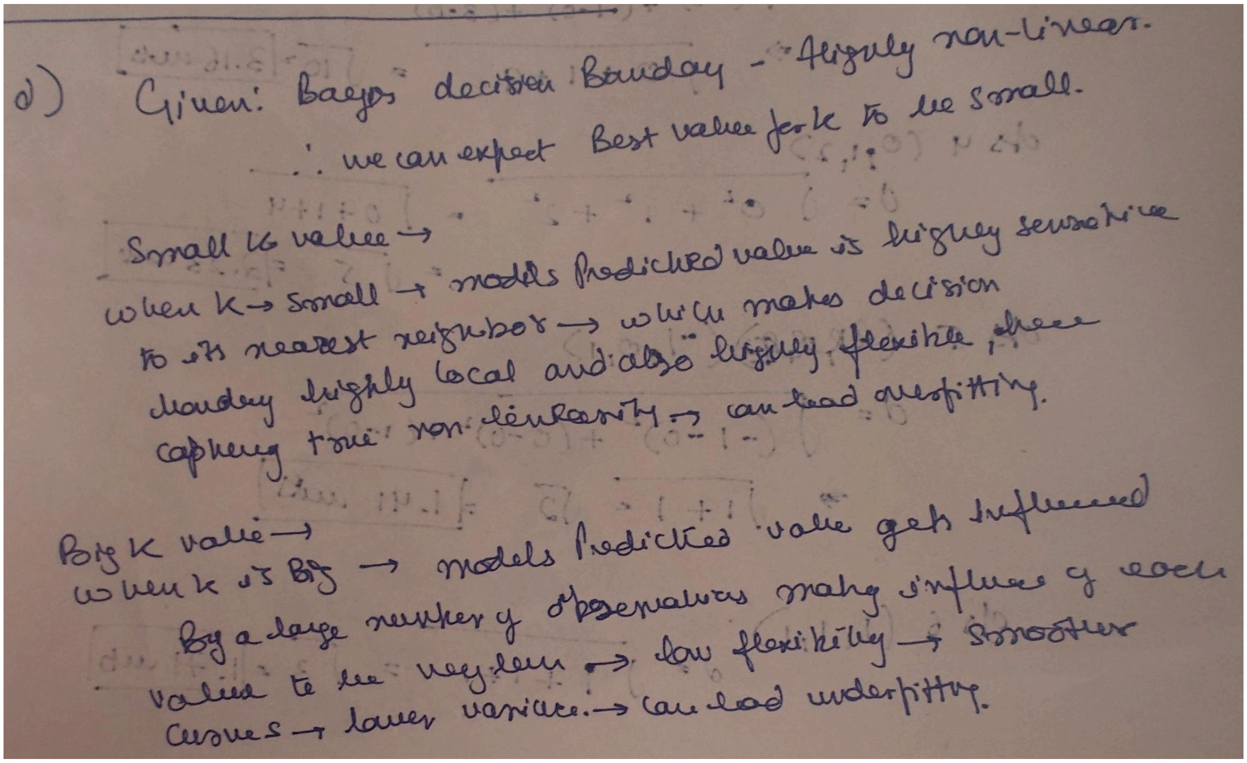
the top 3 nearest values will be taken into consideration.

as -

5	1.41	Green	] 1 vote
6	1.71	Red	] 2 votes
2	2.0	Red	

] - here according to simple principle of majority the resulting response variable $\boxed{Y = \text{Red}}$

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?



d) Given: Bayes decision Boundary - Highly non-linear.

∴ we can expect Best value for k to be small.

Small k value →

when k → small → models predicted value is highly sensitive to its nearest neighbor → which makes decision boundary highly local and also highly flexible, these capturing true non linearity → can lead overfitting.

Big K value →

when k is Big → models predicted value gets influenced By a large number of observations making influence of each value to be very less → low flexibility → smoother curves → lower variance → can lead underfitting.