Unit - 2

Supervised Learning:

- Learning from Observations,
- Bias and Variance,
- Occam’s Razor Principle and Over fitting Avoidance,
- Heuristic Search in Inductive Learning,
- Estimating Generalization Errors ,

Statistical Learning:

- Machine Learning and Inferential Statistical Analysis,
- Descriptive Statistics in Learning Techniques
  - Representing Uncertainties in Data: Probability Distributions
  - Descriptive Measures of Probability Distributions
  - Data Similarity

What is supervised learning?

- Supervised learning is a learning in which we train the machine using data which is well labelled that means some data is already tagged with the correct answer.
- After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

Example:

- Suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

(i) If shape of object is rounded and depression at top having color Red then it will be labelled as –Apple.

(ii) If shape of object is long curving cylinder having color Green-Yellow then it will be labelled as – Banana.

- Now suppose after training the data, you have given a new separate fruit say apple from basket and asked to identify it.

- Since the machine has already learned the things from previous data and this time have to use it wisely.
- It will first classify the fruit with its shape and color and would confirm the fruit name as apple and put it in Apple category.
- Thus the machine learns the things from training data (basket containing fruits) and then apply the knowledge to test data (new fruit).

1. Learning from the observation

What is learning?

- Acquire and organize knowledge, Discover new knowledge, Acquire skills.
- *Learning-from-Observation* is the framework to generate machine's movement to achieve a target task with less user's programming effort.
- In this framework, a user just demonstrates the target task and a machine learns the method to reproduce the target task from the observation.

There is some set S of possible patterns/observations/samples over which various output functions may be defined. The training experience is available in the form of N patterns $s^{(i)} \in S$; $i = 1, 2, \ldots, N$. We specify a pattern by a fixed number n of attributes/features x_j; $j = 1, 2, \ldots, n$; where each feature has real numerical value for the pattern. The domain of x_j is given by a set $V_{x_j} \in \Re$ of its values. A data pattern $s^{(i)}$ has the feature-value set $\{x_1^{(i)}, \ldots, x_n^{(i)}\}$, where $x_j^{(i)} \in V_{x_j}$. We can visualize each pattern with n numerical features as a point in n-dimensional state space $\Re^n$:

$$\mathbf{x} = [x_1\ x_2 \cdots x_n]^T \in \Re^n$$

The set $\mathbf{X}$ is the finite set of feature vectors $\mathbf{x}^{(i)}$ for all the N patterns. We can visualize $\mathbf{X}$ as a region of the state space $\Re^n$ to which the patterns belong, i.e., $\mathbf{X} \subset \Re^n$. Note that $\mathbf{x}^{(i)}$ is the representation of $s^{(i)}$; and $\mathbf{X}$ is the *representation space*.

1.1 Empirical Risk Minimization (ERM)

- It is a principle in statistical learning theory which defines a family of learning algorithms and is used to give theoretical bounds on their performance.
- The idea is that we don’t know exactly how well an algorithm will work in practice (the true "risk") because we don't know the true distribution of data that the algorithm will work on but as an alternative we can measure its performance on a known set of training data.
- We assumed that our samples come from this distribution and use our dataset as an approximation. If we compute the **loss** using the data points in our dataset, it’s called **empirical risk.**

- It is “empirical” and not “true” because we are using a dataset that’s a subset of the whole population.
- When our learning model is built, we have to pick a function that minimizes the empirical risk that is the delta between predicted output and actual output for data points in the dataset.
- This process of finding this function is called empirical risk minimization (ERM). We want to minimize the true risk.
- We don’t have information that allows us to achieve that, so we hope that this empirical risk will almost be the same as the true empirical risk.

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)].$$

1.2 Inductive Learning

- Inductive learning is also called “discovery learning”.
- From the perspective of inductive learning, we are giving input samples (x) and output samples (f(x)) and the problem is to estimate the function (f).
- Simply we can say that generalize the rule from specific samples and mapping to be useful to estimate the output for new samples in the future.
  - **Classification**: when the function being learned is discrete.
  - **Regression**: when the function being learned is continuous

Some practical examples of induction are:

- **Credit risk assessment**.
  - The x is the properties of the customer.
  - The f(x) is credit approved or not.
- **Disease diagnosis**.
  - The x are the properties of the patient.
  - The f(x) is the disease they suffer from.
- **Face recognition**.
  - The x are bitmaps of peoples faces.
  - The f(x) is to assign a name to the face.
- **Automatic steering**.
  - The x are bitmap images from a camera in front of the car.
  - The f(x) is the degree the steering wheel should be turned.

When Should You Use Inductive Learning?

4 problems where inductive learning might be a good idea:

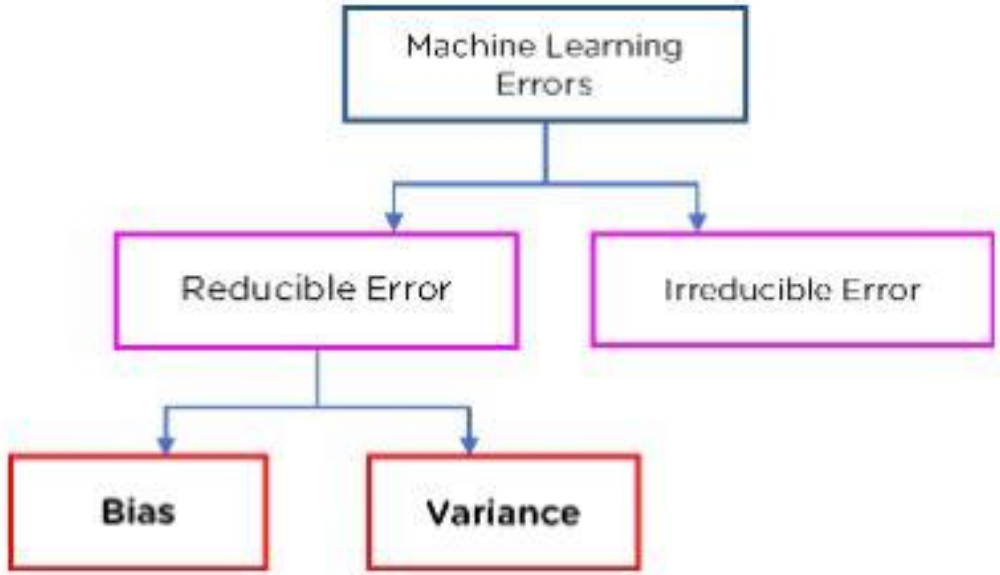
- **Problems where there is no human expert**. If people do not know the answer they cannot write a program to solve it. These are areas of true discovery.
- **Humans can perform the task but no one can describe how to do it**. There are problems where humans can do things that computer cannot do or do well. Examples include riding a bike or driving a car.
- **Problems where the desired function changes frequently**. Humans could describe it and they could write a program to do it, but the problem changes too often. It is not cost effective. Examples include the stock market.
- **Problems where each user needs a custom function**. It is not cost effective to write a custom program for each user. Example is recommendations of movies or books on Netflix or Amazon

2. Bias and Variance

What is an Error?

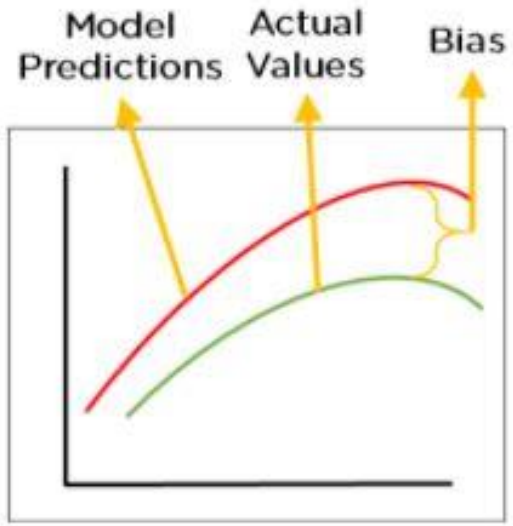
-> an error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset.

-> There are mainly two types of errors in machine learning, which are:



What is Bias?

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.



- A model can have:
  - Low Bias: model will make fewer assumptions about the form of the target function.
  - High Bias: A high bias model also cannot perform well on new data (underfitting).

- Some examples of machine learning algorithms with low bias are Decision Trees, k-Nearest Neighbors and Support Vector Machines. At the same time, an algorithm with high bias is Linear Regression, Linear Discriminant Analysis and Logistic Regression.

Ways to reduce High Bias:

High bias mainly occurs due to a much simple model. Below are some ways to reduce the high bias:

- Increase the input features as the model is underfitted.
- Decrease the regularization term.
- Use more complex models, such as including some polynomial features

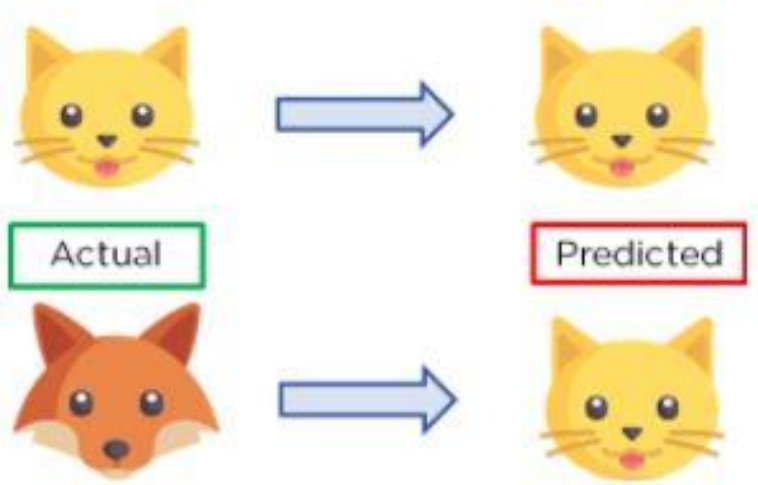
What is Variance?

- variance tells that how much a random variable is different from its expected value.
- Variance errors are either of **low variance or high variance.**
  - **Low variance** means there is a small variation in the prediction of the target function with changes in the training data set.
  - **High variance** shows a large variation in the prediction of the target function with changes in the training dataset
- Since, with high variance, the model learns too much from the dataset, it leads to overfitting of the model.
- The formula for variance is:

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

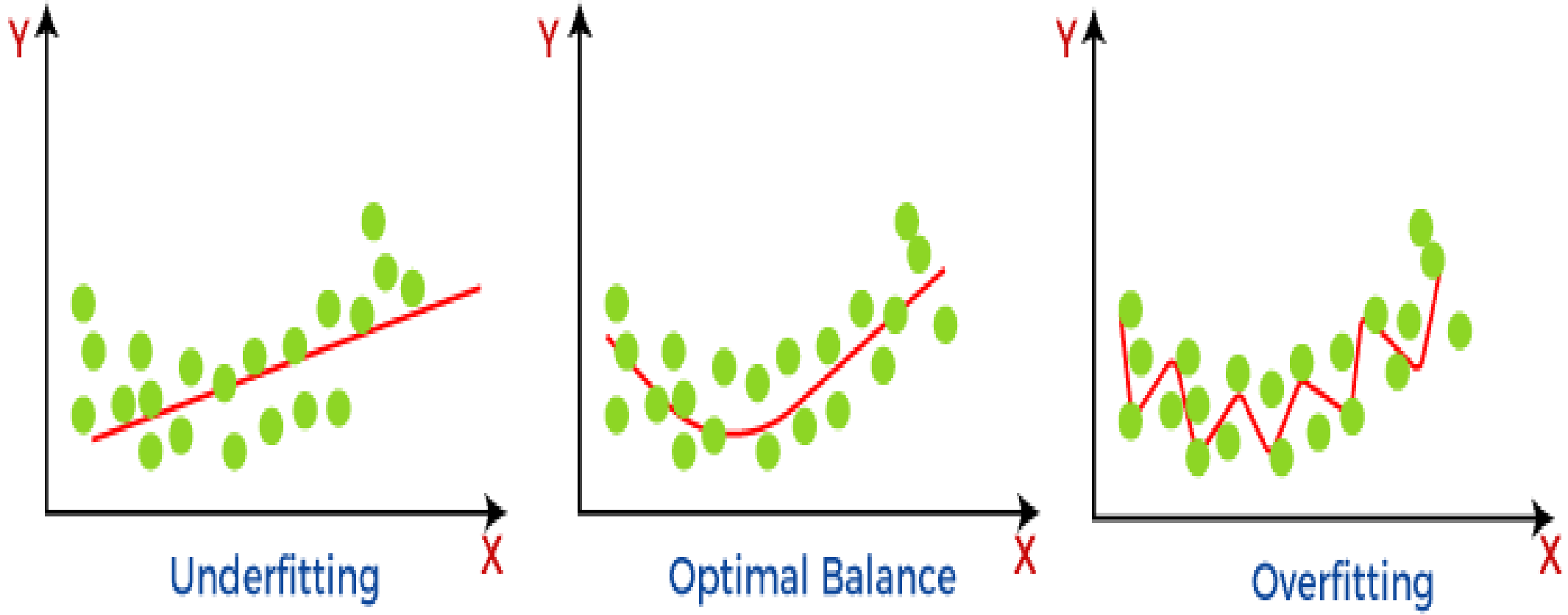
  - Irreducible error (noise in our data)

Example:



- we can see that our model has learned extremely well for our training data, which has taught it to identify cats. But when given new data, such as the picture of a fox, our model predicts it as a cat, as that is what it has learned.
- This happens when the Variance is high, our model will capture all the features of the data given to it, including the noise, will tune itself to the data, and predict it very well but when given new data, it cannot predict on it as it is too specific to training data.

- ***Variance*** refers to the amount of variability or inconsistency in the model's predictions when trained on different subsets of the training data.
- It quantifies how much the model's predictions change based on the specific data samples used for training.



Example on High Bias & Low Variance, Low Bias and High Variance

High Bias, Low Variance (Underfitting):

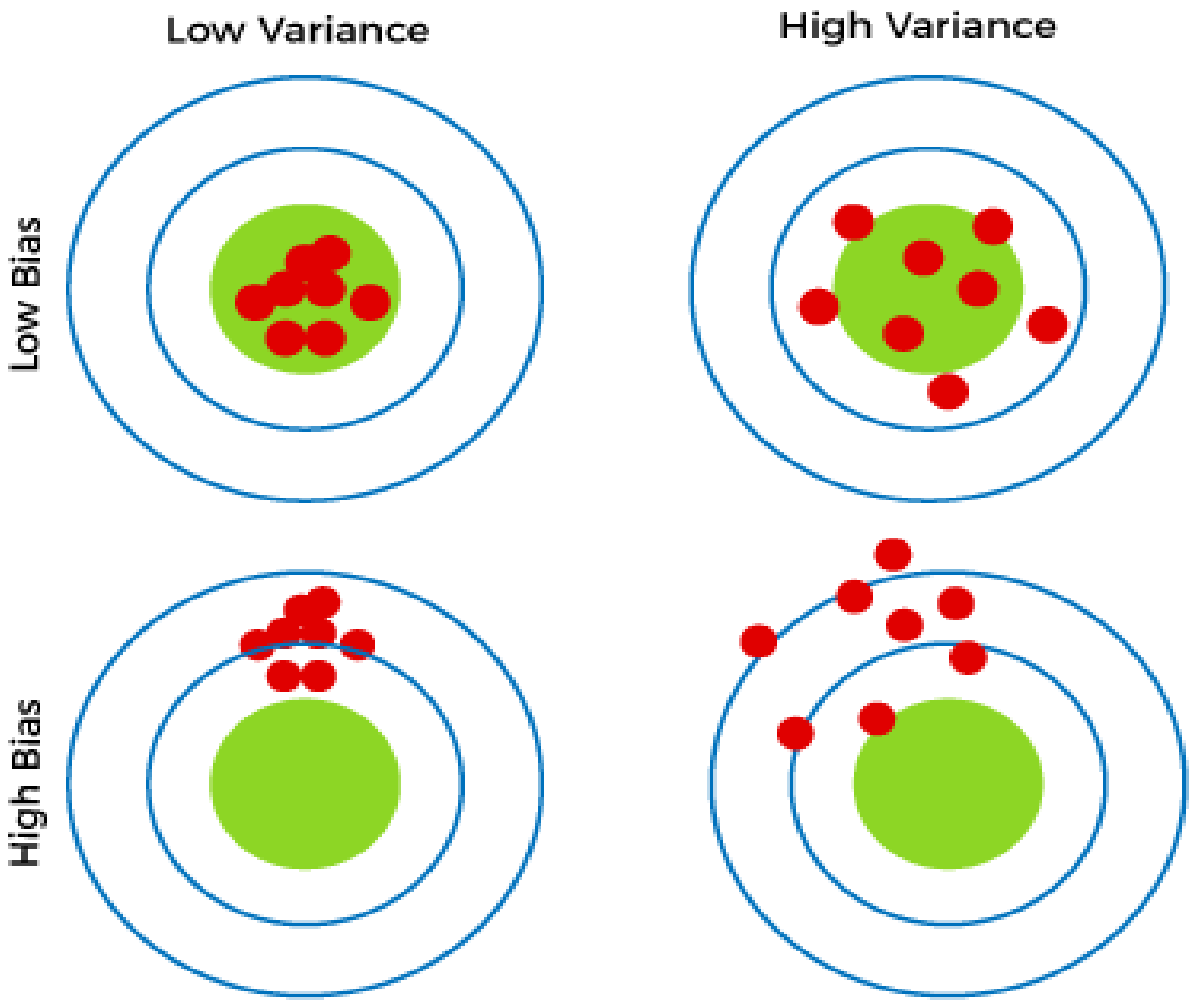
- Suppose we have a dataset of house prices, where the features include the "**Sqrft**" and " **number of bedrooms**".
- If we use a linear regression model with only one feature (e.g: Sqrft), the model might not be able to capture the complexity of the relationship between the features and the house prices.
- It would result in a high bias, as the model is too simplistic to make accurate predictions.
- This would lead to *underfitting*, and the model would perform poorly on both the training data and new, unseen data.

Example on High Bias & Low Variance, Low Bias and High Variance

Low Bias, High Variance (Overfitting):

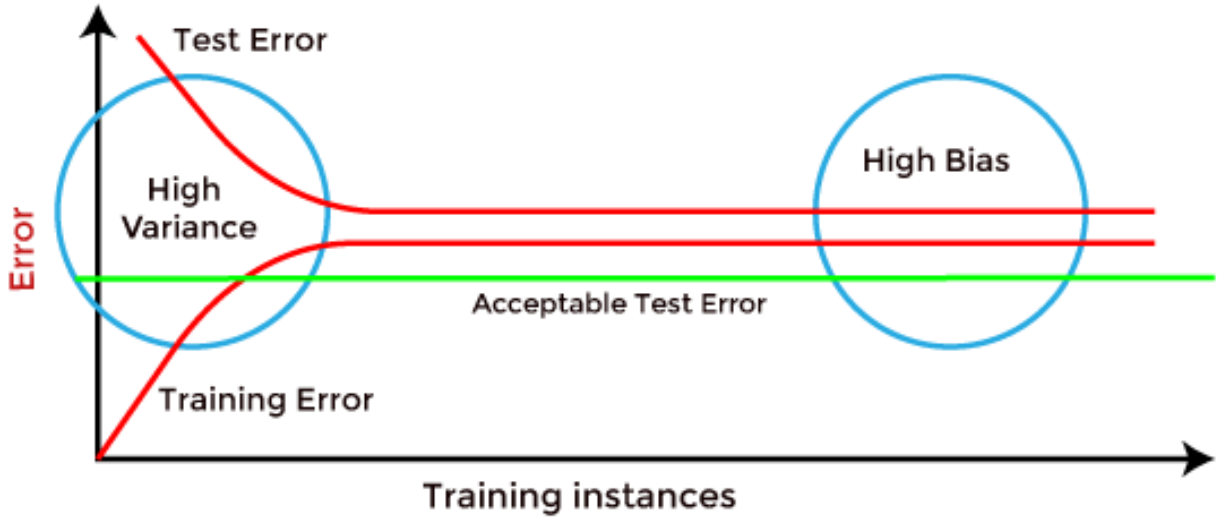
- If we use a very complex model, such as a high-degree polynomial regression, the model may fit the training data extremely well.
- However, this model might also pick up the noise in the training data, effectively memorizing the individual data points.
- This would result in a low bias, as the model can perfectly fit the training data, but a high variance, as it becomes too sensitive to changes in the training data.
- Consequently, the model may perform very well on the training data but generalize poorly to new, unseen data, leading to overfitting

Bias – Variance tradeoff



How to identify High variance and High bias

- High variance can be identified if the model has:
  - Low training error and high test error.
- High Bias can be identified if the model has:
  - High training error and the test error is almost similar to training error.



3. Occam’s Razor Principle

- Occam’s Razor principle is given by the Willam of Occam, was born in 1280. His name is linked with machine learning through the basic idea ‘**The simpler explanations are more reasonable, and any unnecessary complexity should be shaved off** ’.
- If there are two algorithms and both of them perform equally well on the training set, then according to Occam’s razor principle, the simpler algorithm can be expected to do better on a test set;
- ‘simpler’ may imply needing lesser parameters, lesser training time, fewer attributes for data representation, and lesser computational complexity.
- Occam’s razor principle suggests hypothesis functions that avoid overfitting of the training data.

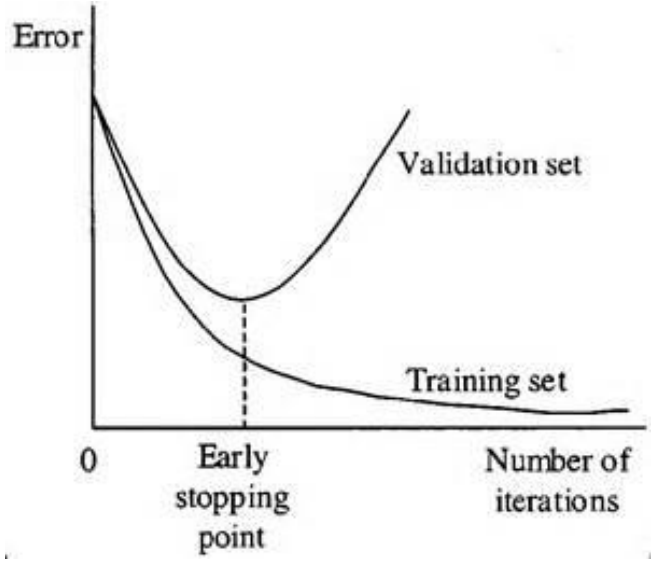
4. Overfitting Avoidance

What is overfitting?

- A statistical model is said to be over fitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.
- And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

Techniques to reduce overfitting:

- Cross validation
- Train the model with more data
- Remove Features
- Ensemble methods: Bagging, Boosting
- Regularization:
- Early stopping: stopping the training process before the learner passes that point.



Regularization

- Regularization is a method used in machine learning to prevent models from memorizing noise and irrelevant patterns in the training data.
- It involves adding a penalty term to the model's objective function during training to discourage the model from overfitting the training data.
- Regularization helps the model generalize better to new, unseen data and improves its overall performance.
- Regularization techniques like L1 (Lasso) and L2 (Ridge) regularization are used.

5. Heuristic Search in Inductive Learning

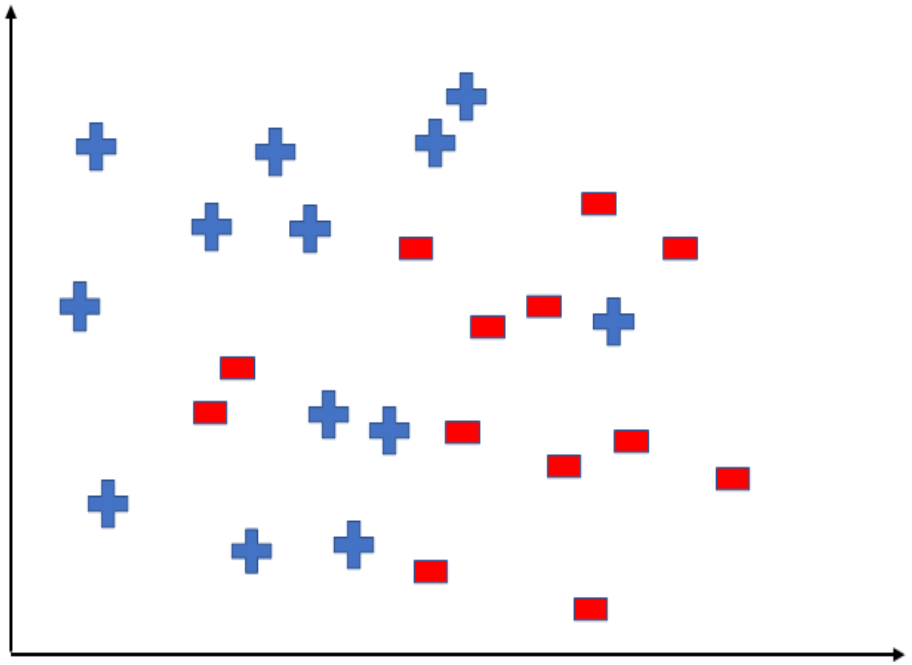
“Heuristic search is class of method which is used in order to search a solution space for an optimal solution for a problem”.

5.1 Search Through Hypothesis Space:

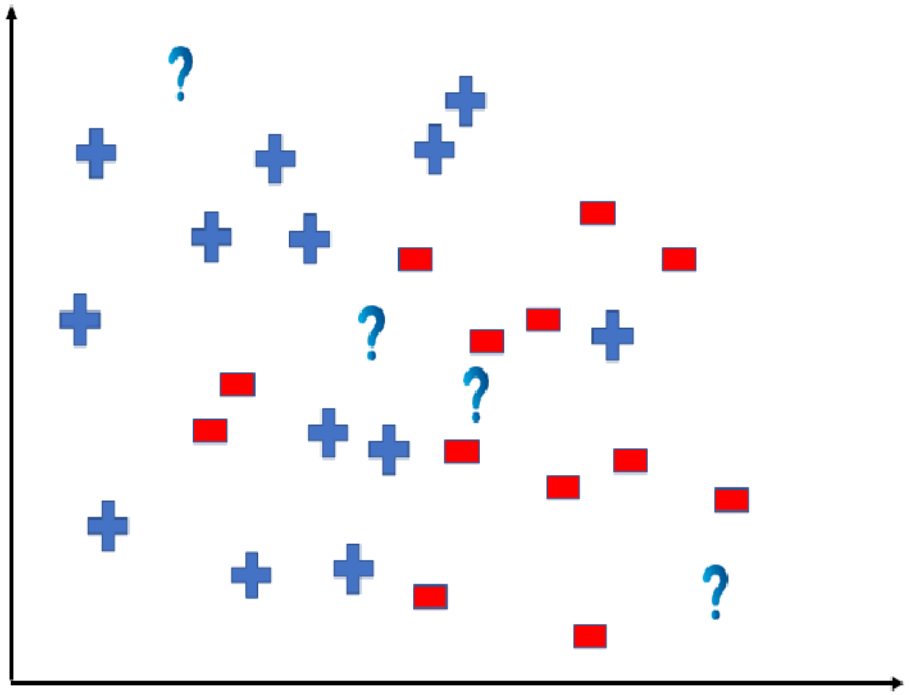
What is hypothesis (h) ?

- Hypothesis is a function that best describes the target in supervised machine learning.
- The hypothesis that an algorithm would come up depends upon the data and also depends upon the restrictions and bias that we have imposed on the data.

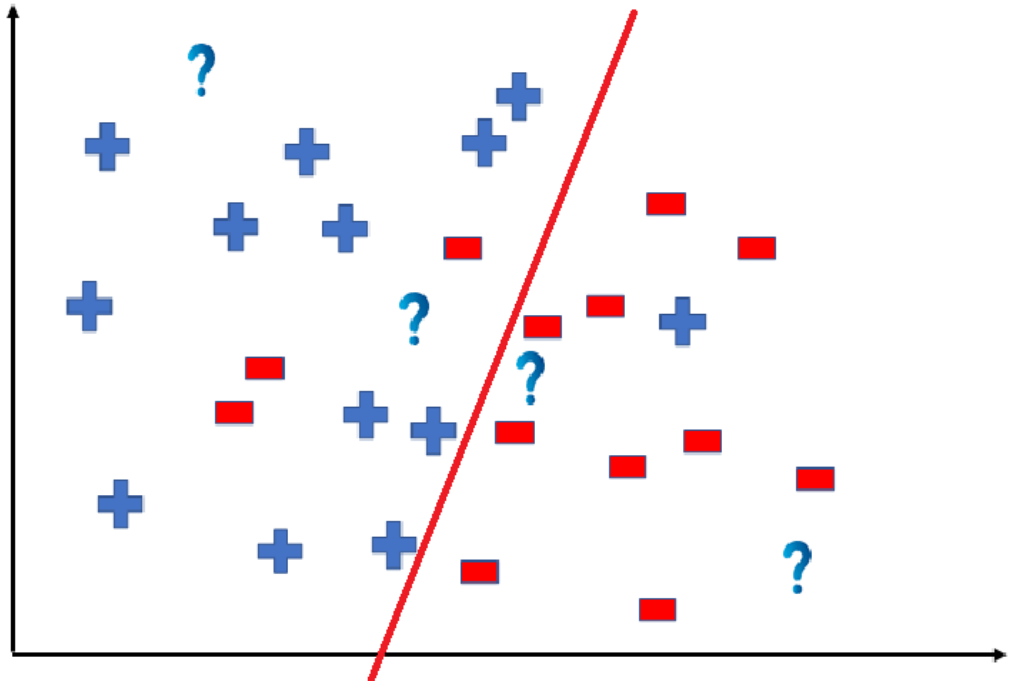
- To better understand the Hypothesis Space and Hypothesis consider the following coordinate that shows the distribution of some data:



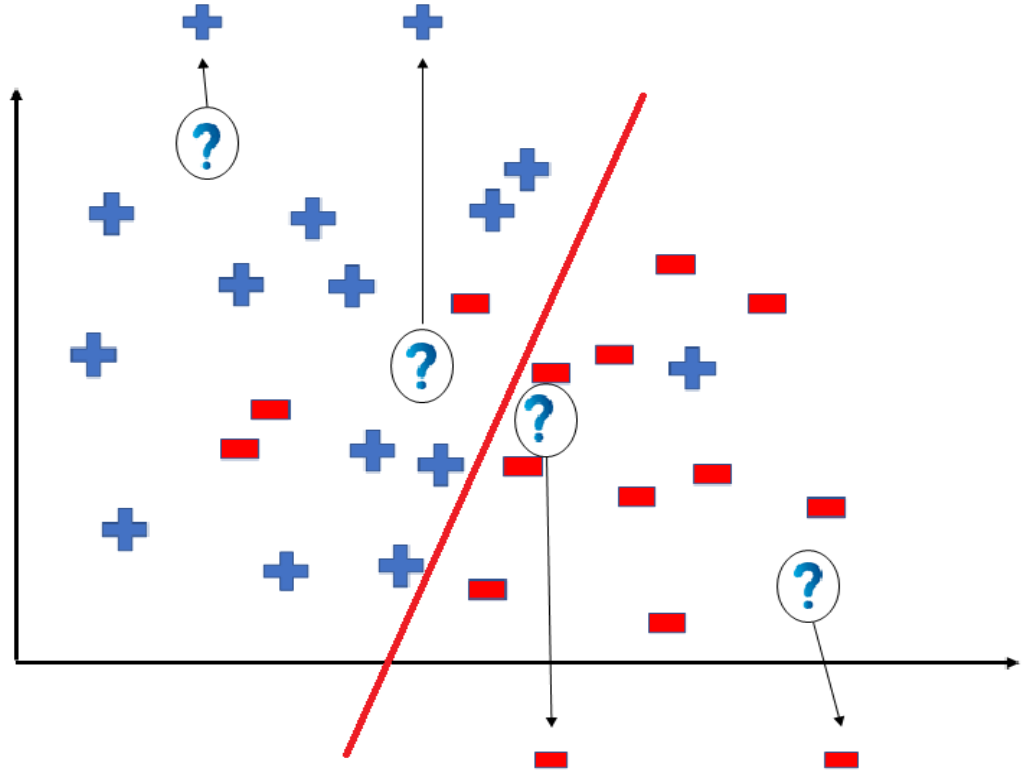
Training data



Test data

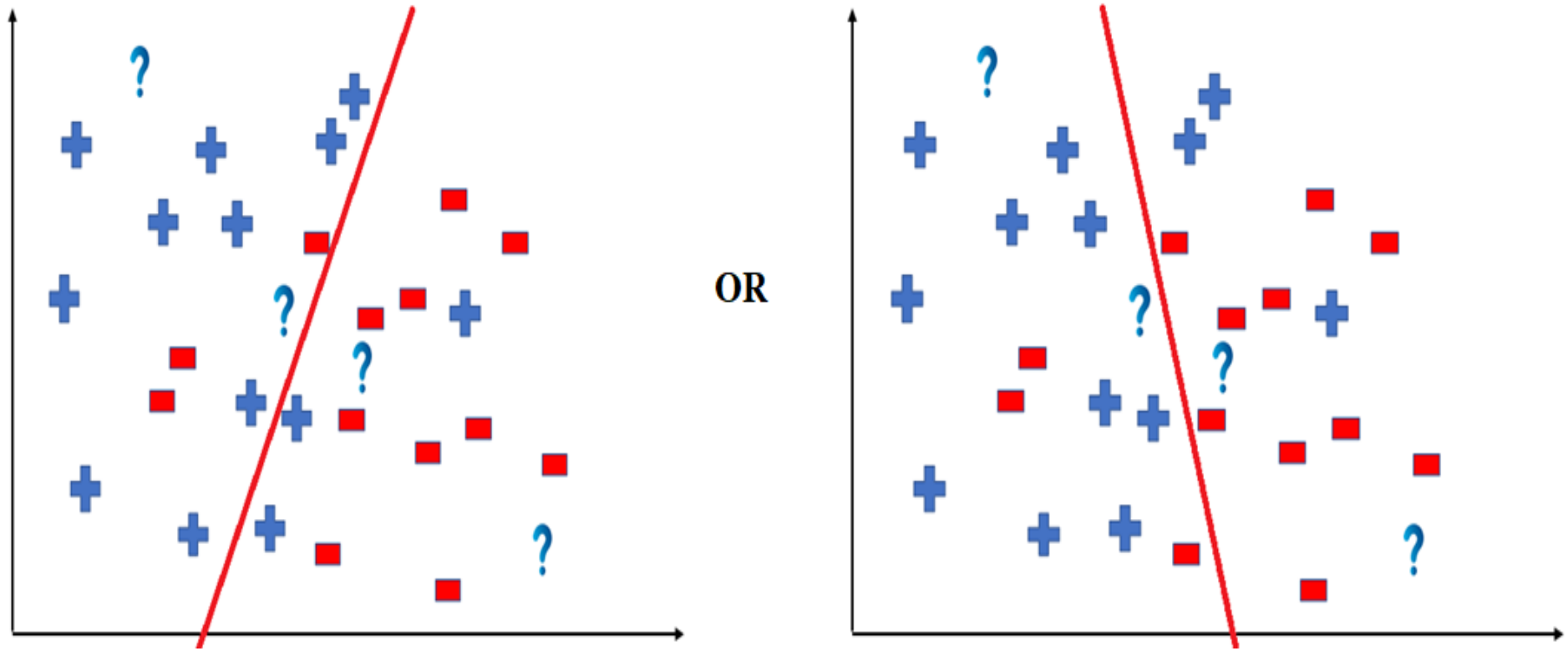


After applying the algorithm we got this plane



classification

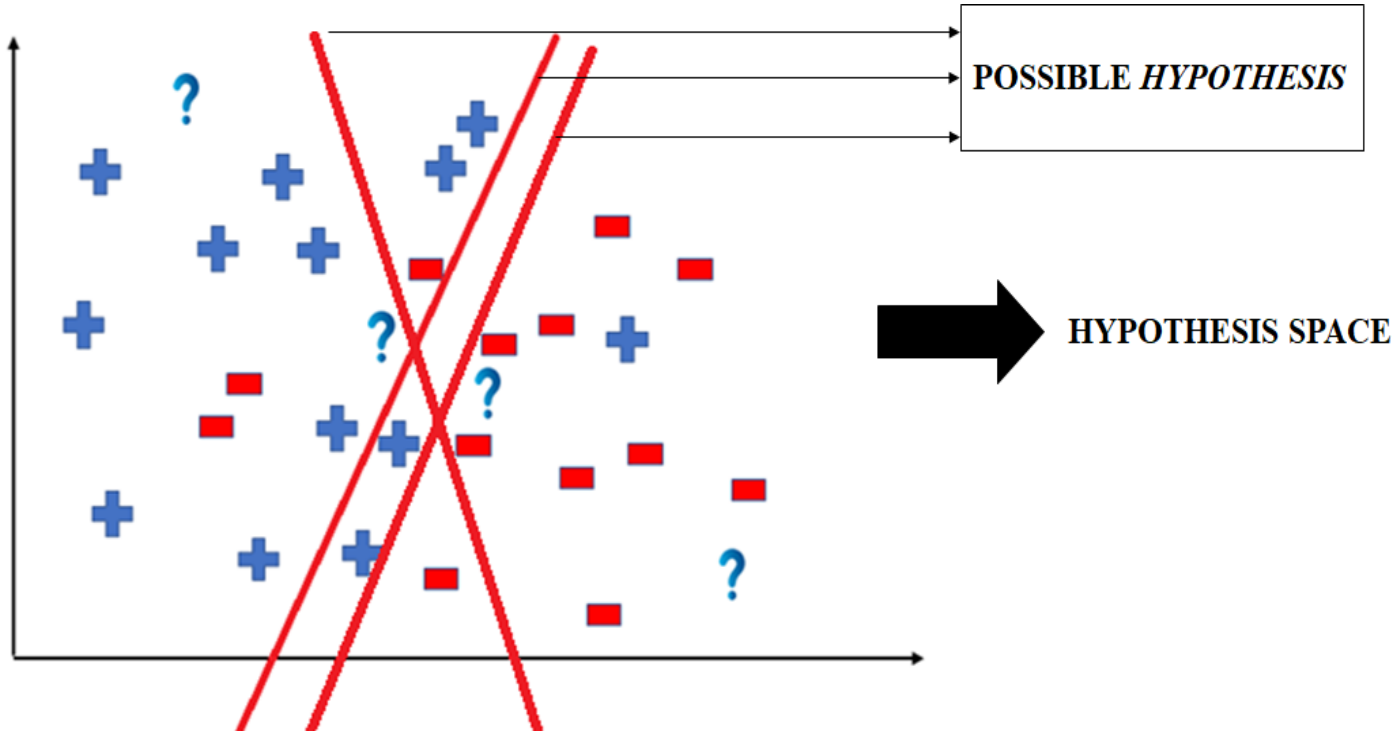
But note here that we could have divided the coordinate plane as:



The way in which the coordinate would be divided depends on the data, algorithm and constraints.

- All these legal possible ways in which we can divide the coordinate plane to predict the outcome of the test data composes of the **Hypothesis Space**.
- Each individual possible way is known as the **hypothesis**.

Hence, in this example the hypothesis space would be like:



- Applied machine learning organizes the search as per the following two-step procedure:
  1. The search is first focused on a class of the possible hypotheses, chosen for the learning task in hand. Prior knowledge and experience are helpful in this selection.
  2. For each of the classes, the corresponding learning algorithm organizes the search through all possible structures of the learning machine.
- In this few techniques used in heuristic search to optimize hypothesis complexity for a given training dataset.
  - Regularization
  - Early stopping
  - Pruning

Regularization:

- It is a technique to prevent the model from overfitting by adding extra information to it.

How does Regularization Work?

Regularization works by adding a penalty or complexity term to the complex model.

Let's consider the simple linear regression equation:

$$y= w0+w1.x1+w2.x2+w3.x3+\cdots+wn.xn +b$$

- In the above equation, Y represents the value to be predicted
- X1, X2, …Xn are the features for Y.
- w0,w1,…..wn are the weights or magnitude attached to the features, respectively. Here represents the bias of the model, and b represents the intercept.

- Formally, it is possible to write the new criterion as a sum of the error on the training set plus a *regularization term*, which depicts constraints after properties of solutions:

$$\bar{E} = E + \lambda \Omega$$

$$= \text{error on data} + \lambda \times \text{hypothesis complexity where } \lambda \text{ gives the weight of penalty}$$

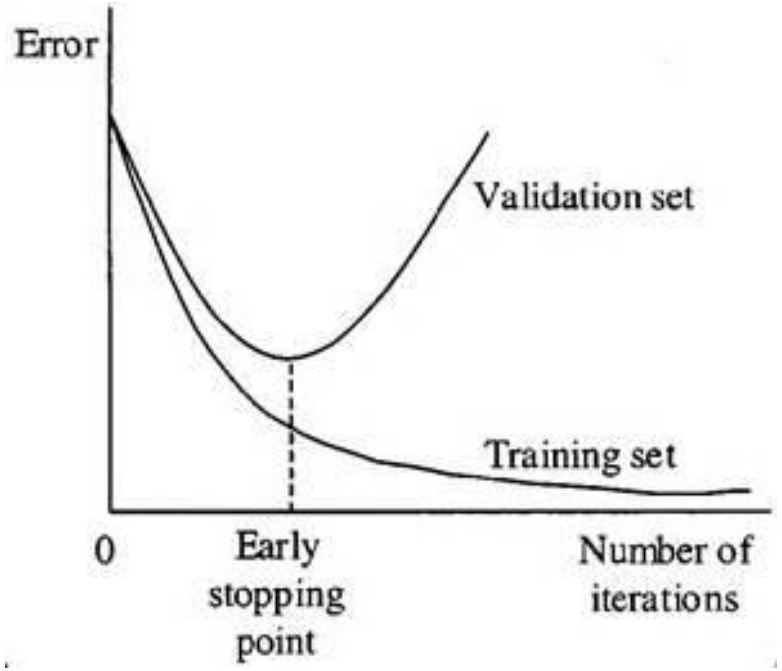
The hypothesis complexity may be expressed as,

$$\Omega = \frac{1}{2} \sum_{l,j} w_{lj}^2$$

- The regularizer encourages smaller weights wij. For small values of weights, the network mapping is approximately linear. Relatively large values of weights lead to overfitted mapping with regions of large curvature

Early Stopping

- **Early stopping** is an optimization technique used to reduce overfitting without compromising on model accuracy.
- The main idea behind early stopping is to stop training before a model starts to overfit.



Pruning

- In general pruning is a process of removal of selected part of plant such as bud, branches and roots .
- In Decision Tree pruning does the same task it removes the branches of decision tree to overcome the overfitting condition of decision tree.
- There are two approaches of pruning:
  - Pre – pruning : Stop growing the tree before it reaches perfection
  - Post - pruning: allow the tree to grow entirely and then post – prune some of the branches from it.

5.2 Ensemble Learning

- **The ensemble methods in machine learning combine the insights obtained from multiple learning models to facilitate accurate and improved decisions.**
- In learning models, noise, variance, and bias are the major sources of error.
- The ensemble methods in machine learning help minimize these error-causing factors, thereby ensuring the accuracy and stability of machine learning (ML) algorithms.

Example

- If you are planning to buy an **air-conditioner**, would you enter a showroom and buy the air-conditioner that the salesperson shows you? The answer is probably no.
- In this day and age, you are likely to ask your friends, family, and colleagues for an opinion, do research on various portals about different models, and visit a few review sites before making a purchase decision.
- In a nutshell, you would not come to a conclusion directly. Instead, you would try to make a more informed decision after considering diverse opinions and reviews.

Ensemble Techniques (basic)

Mode:

- In statistical terminology, "mode" is the number or value that most often appears in a dataset of numbers or values.
- In this ensemble technique, machine learning professionals use a number of models for making predictions about each data point.
- The predictions made by different models are taken as separate votes.
- Subsequently, the prediction made by most models is treated as the ultimate prediction.

Ensemble Techniques (basic)

The Mean/Average:

- In the mean/average ensemble technique, data analysts take the average predictions made by all models into account when making the ultimate prediction.

Example:

- Let's take, for instance, one hundred people rated the beta release of your travel and tourism app on a scale of 1 to 5,
- where 15 people gave a rating of 1, 28 people gave a rating of 2, 37 people gave a rating of 3, 12 people gave a rating of 4, and 8 people gave a rating of 5.
- The average in this case is - (1 * 15) + (2 * 28) + (3 * 37) + (4 * 12) + (5 * 8) / 100 = 2.7

Ensemble Techniques (basic)

The Weighted Average:

- In the weighted average ensemble method, data scientists assign different weights to all the models in order to make a prediction, where the assigned weight defines the relevance of each model.
- As an example, let's assume that out of 100 people who gave feedback for your travel app, 70 are professional app developers, while the other 30 have no experience in app development.
- In this scenario, the weighted average ensemble technique will give more weight to the feedback of app developers compared to others.

Ensemble Techniques (advanced)

Bagging (Bootstrap Aggregating):

- The primary goal of "bagging" or "bootstrap aggregating" ensemble method is to minimize variance errors in decision trees.
- The objective here is to randomly create samples of training datasets with replacement (subsets of the training data).
- The subsets are then used for training decision trees or models. Consequently, there is a combination of multiple models, which reduces variance, as the average prediction generated from different models is much more reliable and robust than a single model or a decision tree.

Boosting:

- An iterative ensemble technique, "boosting," adjusts an observation's weight based on its last classification.
- In case observation is incorrectly classified, "boosting" increases the observation's weight, and vice versa. Boosting algorithms reduce bias errors and produce superior predictive models.

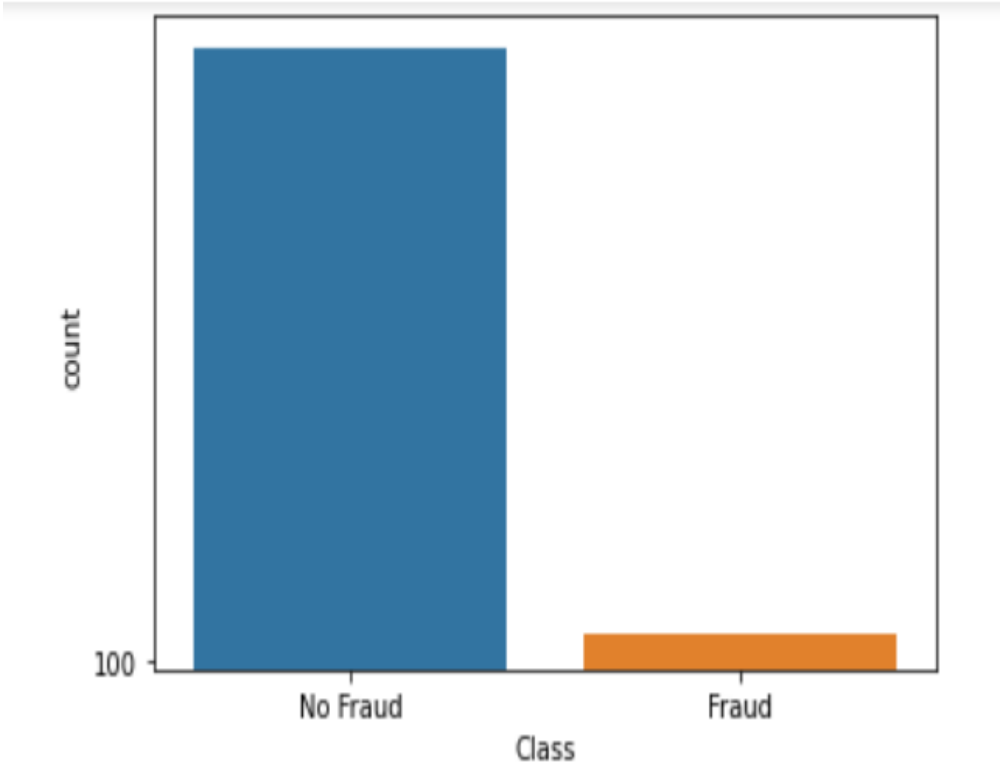
- In the boosting ensemble method, data scientists train the first boosting algorithm on an entire dataset and then build subsequent algorithms by fitting residuals from the first boosting algorithm, thereby giving more weight to observations that the previous model predicted inaccurately.

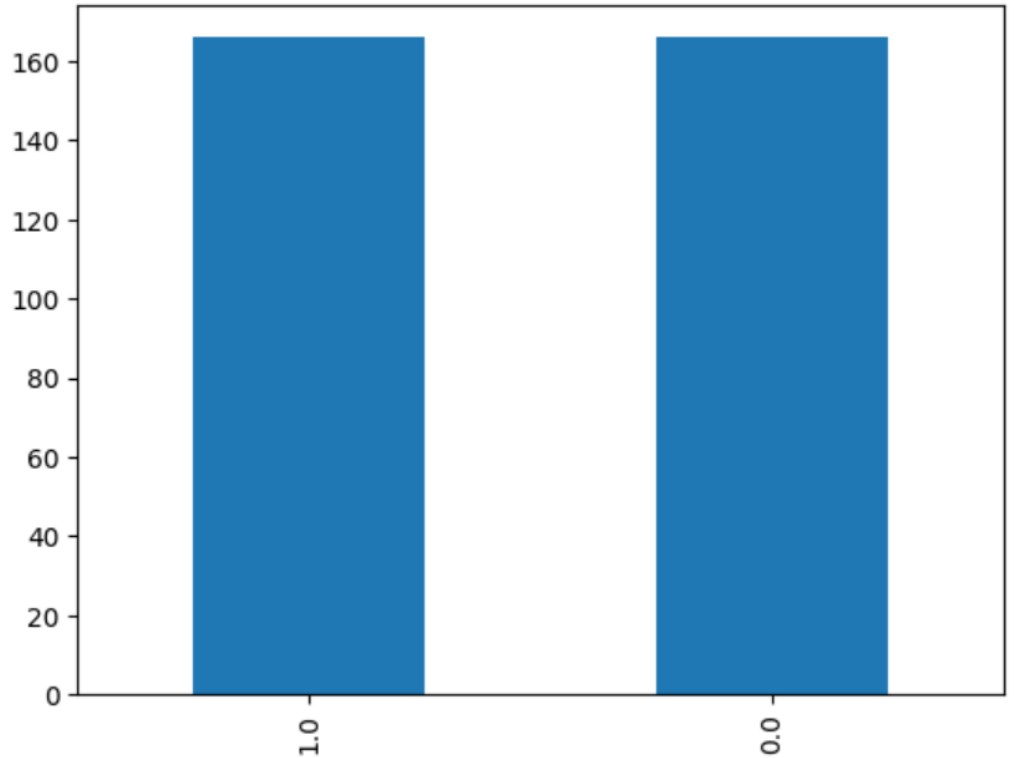
5.2.1 Class imbalance problems

- Ensemble methods have been used to solve **class-imbalanced problems.**
- Class Imbalance is a common problem in machine learning, especially in classification problems. Imbalance data can hamper our model accuracy big time. It appears in many domains, including fraud detection, spam filtering, disease screening etc
- **Two-class / multiclass-imbalanced** data are class-imbalanced if the primary class of interest is represented by just a few samples in the dataset, whereas the majority of the samples represent the other class.
- The general approaches for improvement of the classification performance of class-imbalanced data are: (i) Oversampling (ii) Undersampling (iii) Threshold moving (iv) Ensemble methods

Example :

- Credit Card Fraud Detection Example





5.3 Evaluation of a Learning System

- **Accuracy**: Based on the performance of a model.
- **Robustness:** ‘Robustness’ means that the machine can perform adequately under all circumstances including the cases when information is corrupted by noise, is incomplete, and is interfered with irrelevant data.
- **Computational Complexity and Speed**: Computational complexity of a learning algorithm and learning speed determine the efficiency of a learning system: how fast the systems can arrive at a correct answer, and how much computer memory is required. We know how important speed is in real-time situations.

5.3 Evaluation of a Learning System (cont.)

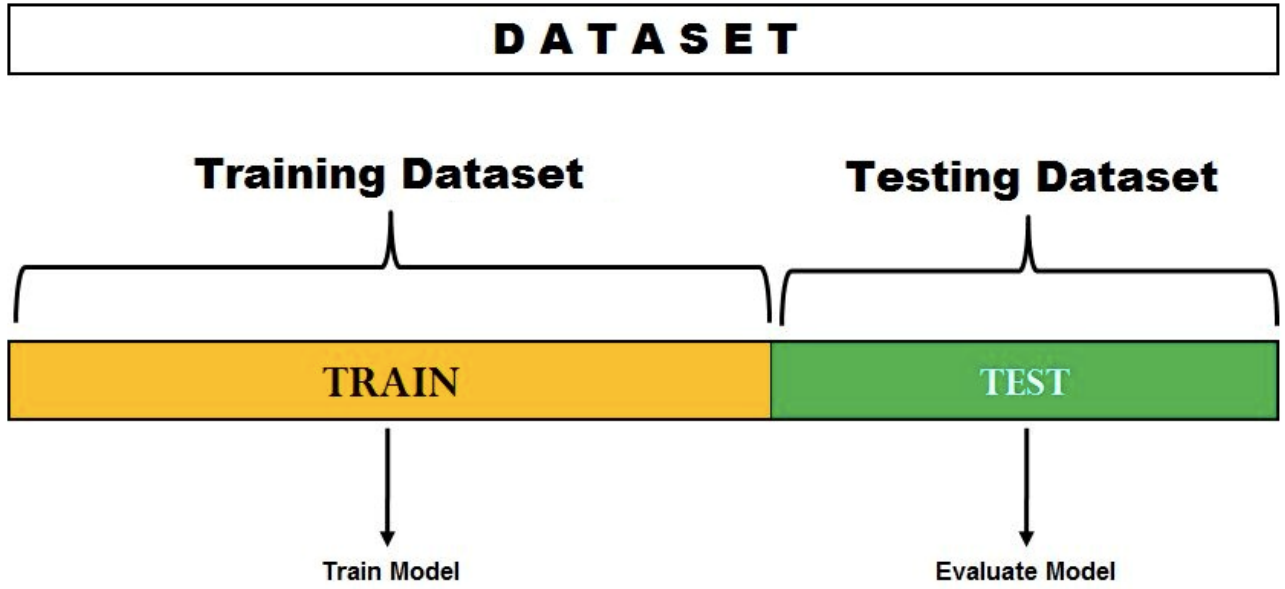
- **Online Learning:** An online learning system can continue to assimilate new data. This feature is essential for a learning system which continues to receive inputs from a real-time environment.
- **Scalability**: This is the capability to build the learning machine considering the huge amounts of data. Typically, the assessment of scalability is done with a series of datasets of ascending size
- **Interpretability**: This is the level of understanding and insight offered by a learning algorithm.

6. Estimating Generalization Errors

- Generalization assesses a model’s ability to process new data and generate accurate predictions after being trained on a training set.
- There are different techniques involved to estimate generalized errors
  - Holdout method
  - Random sampling
  - Cross validation
  - Bootstrapping

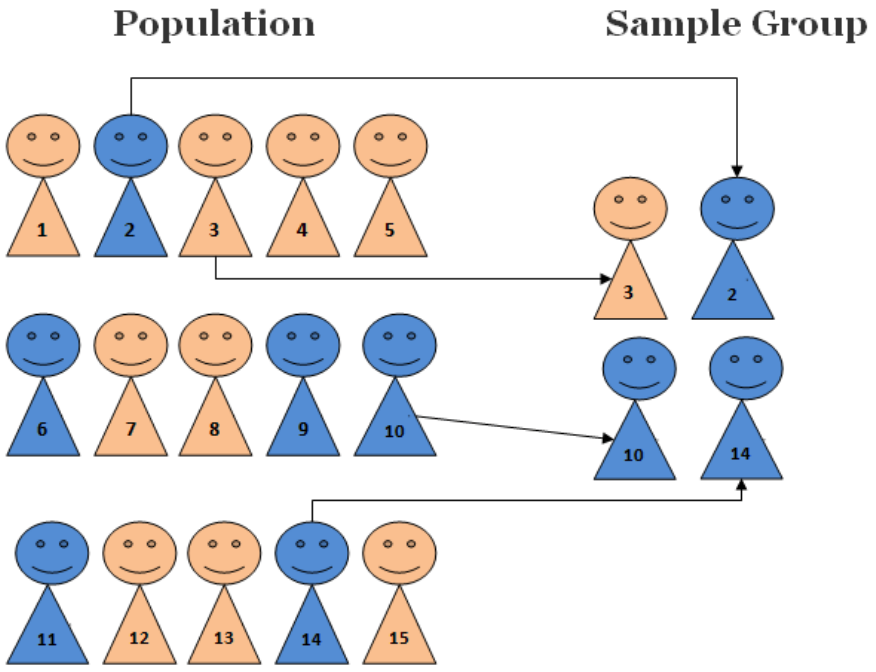
Holdout method

- Hold-out is when you split up your dataset into a ‘train’ and ‘test’ set.
- The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data.
- A common split when using the hold-out method is using 80% of data for training and the remaining 20% of the data for testing.



Random Sampling

- Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen.
- A sample chosen randomly is meant to be an unbiased representation of the total population.
- If for some reasons, the sample does not represent the population, the variation is called a sampling error.



How to Conduct a Random Sample

Step 1: Define the Population

Example: I wish to learn how the stocks of the largest companies in the United States have performed over the past 20 years. My population is the largest companies in the United States as determined by the S&P 500.

Step 2: Choose Sample Size

Example: My sample size will be 20 companies from the S&P 500.

Step 3: Determine Population Units

Example: Using exchange information, I copy the companies comprising the S&P 500 into an Excel spreadsheet

Step 4: Assign Numerical Values

Example: I assign the numbers 1 through 500 to the companies in the S&P 500 based on alphabetical order of the current CEO, with the first company receiving the value '1' and the last company receiving the value '500'.

Step 5: Select Random Values

Example: Using the random number table, I select the numbers 2, 7, 17, 67, 68, 75, 77, 87, 92, 101, 145, 201, 222, 232, 311, 333, 376, 401, 478, and 489.

Step 6: Identify Sample

Example: My sample consists of the 2nd item in the list of companies alphabetically listed by CEO's last name. My sample also consists of company number 7, 17, 67, etc.

Cross validation

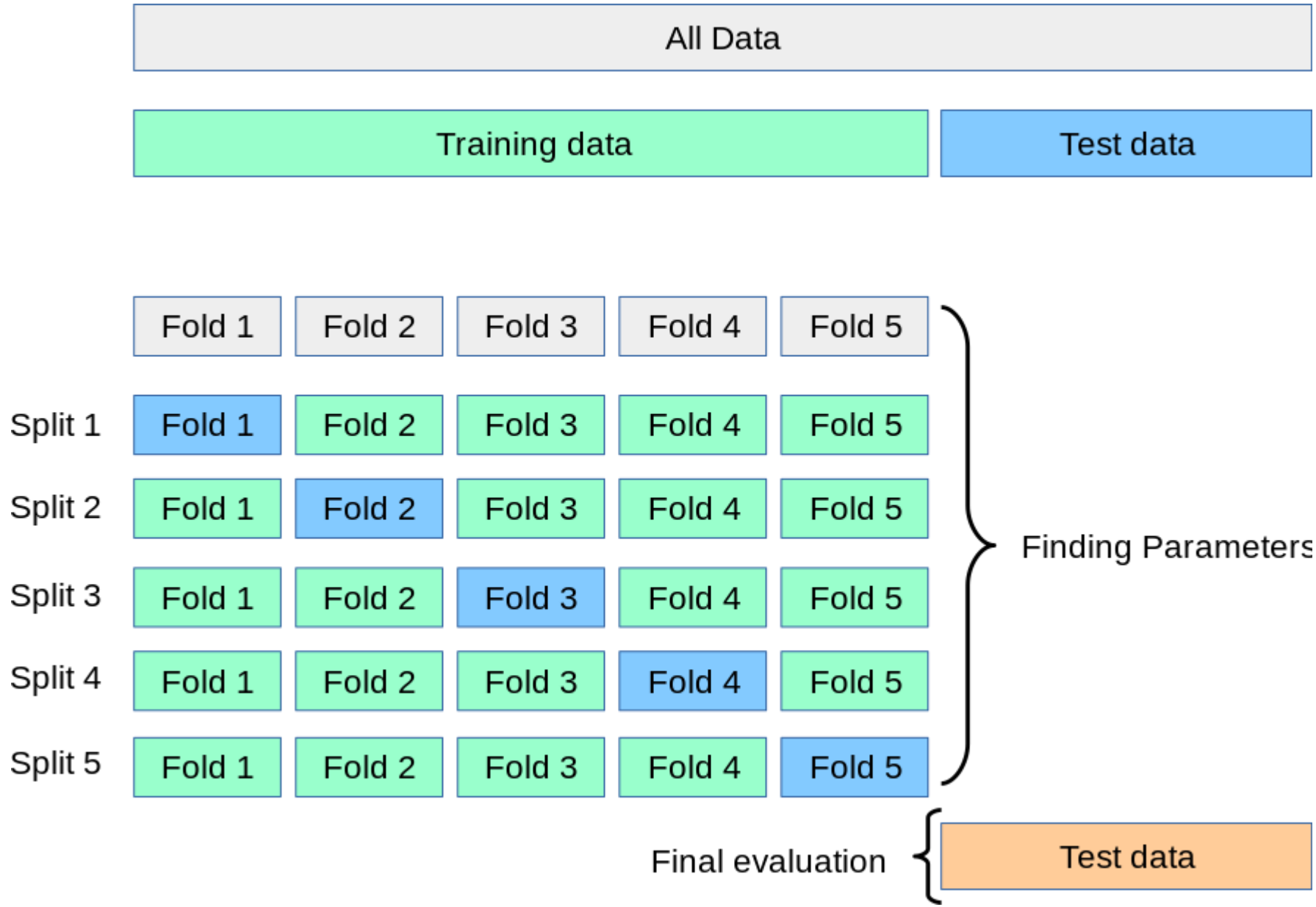
- Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.
- The major goal of cross validation is to train on as much data as possible.

Methods used for Cross – Validation:

- K-fold cross validation
- Leave one out cross validation

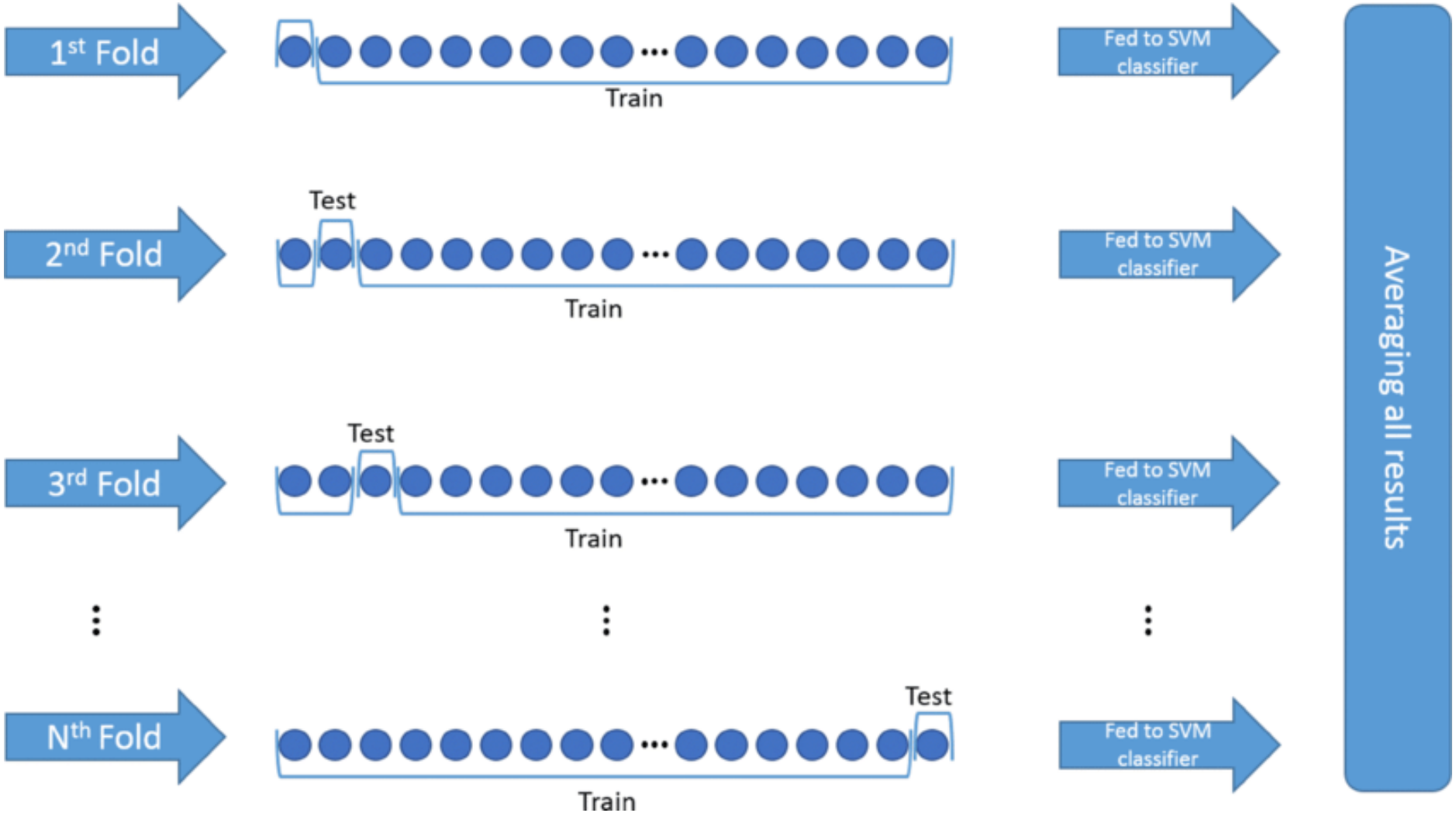
K- fold cross validation

- K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called **folds**.
- For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set.
- The steps for k-fold cross-validation are:
  - Split the input dataset into K groups
  - For each group:
    - Take one group as the reserve or test data set.
    - Use remaining groups as the training dataset
    - Fit the model on the training set and evaluate the performance of the model using the test set.


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters
Final evaluation
Test data

Leave one out cross validation (LOOCV)

- In this method, we divide the data into train and test sets – but with a twist. Instead of dividing the data into 2 subsets, we select a single observation as test data, and everything else is labeled as training data and the model is trained.
- Now the 2nd observation is selected as test data and the model is trained on the remaining data.
- This process continues ‘n’ times and the average of all these iterations is calculated and estimated as the test set error.


1st Fold
Train
Fed to SVM classifier
2nd Fold
Test
Train
Fed to SVM classifier
3rd Fold
Test
Train
Fed to SVM classifier
Nth Fold
Test
Train
Fed to SVM classifier
Averaging all results

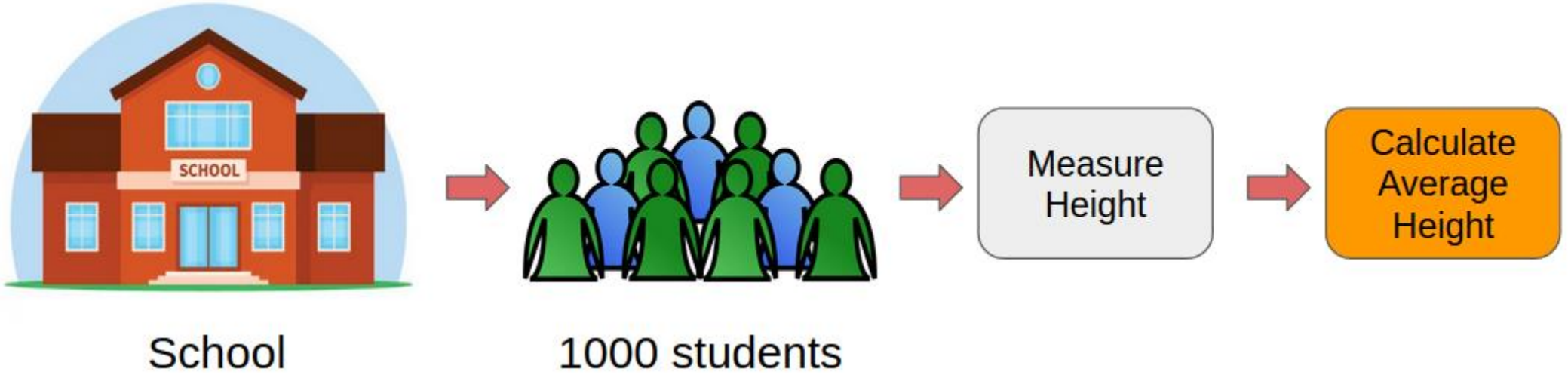
Bootstrapping

In statistics, Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

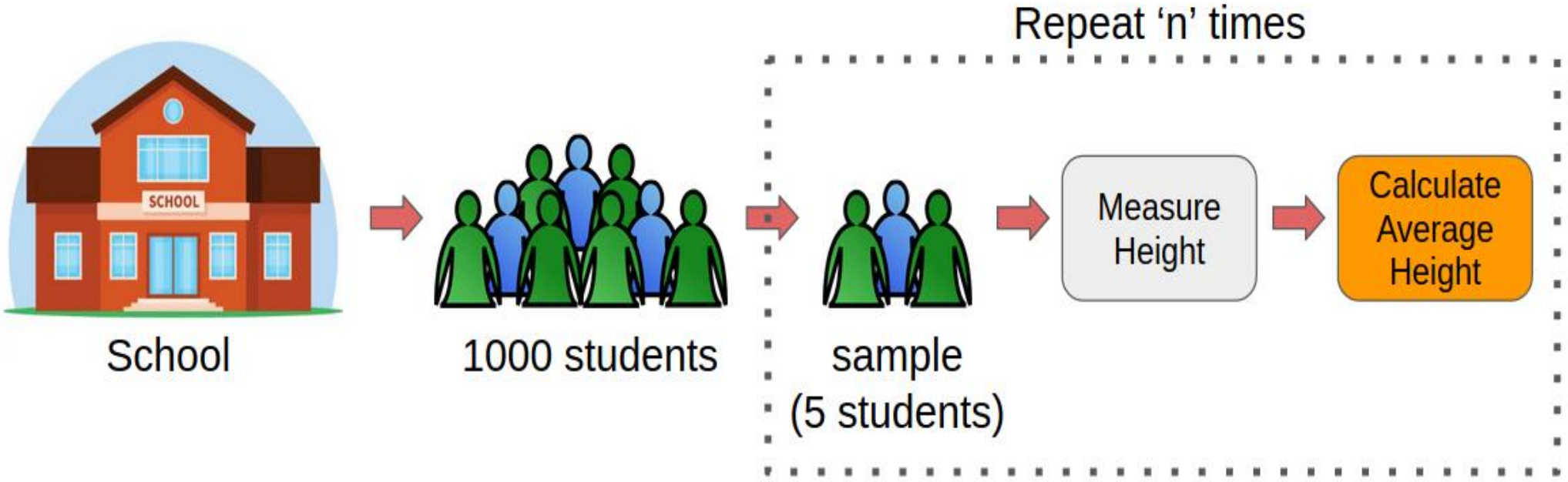
- **Sampling:** With respect to statistics, sampling is the process of selecting a subset of items from a vast collection of items (population) to estimate a certain characteristic of the entire population
- **Sampling with replacement:** It means a data point in a drawn sample can reappear in future drawn samples as well
- **Parameter estimation:** It is a method of estimating parameters for the population using samples.

Why Do We Need Bootstrap Sampling?

- Let’s say we want to find the mean height of all the students in a school (which has a total population of 1,000). So, how can we perform this task?
- One approach is to measure the height of all the students and then compute the mean height



- Instead of measuring the heights of all the students, we can draw a random sample of 5 students and measure their heights. We would repeat this process 20 times and then average the collected height data of 100 students (5 x 20). This average height would be an estimate of the mean height of all the students of the school.



Statistical Learning

Population and Sample

Before discussing the Inferential statistics, let us see the population and sample.

- Population contains “all the data points from a set of data”.
- It is a group from where we collect the data.
- While a sample consists of “some observations selected from the population”.
- The sample from the population should be selected such that it has all the characteristics that a population has.
- Population’s measurable characteristics such as mean, standard deviation etc. are called as parameters
- while Sample’s measurable characteristic is known as a statistic.

Inferential Statistical Analysis

- In Inferential statistics, we make an inference from a sample about the population.
- The main aim of inferential statistics is to draw some conclusions from the sample and generalize them for the population data. **E.g. we have to find the average salary of a data analyst across India.** There are two options**.**

1.The first option is to consider the data of data analysts across India and ask them their salaries and take an average.

2.The second option is to take a sample of data analysts from the major IT cities in India and take their average and consider that for across India.

Descriptive Statistics In Learning Techniques

- **Descriptive statistics** ”summarize and organize characteristics of a data set”. A data set is a collection of responses or observations from a sample or entire population.
- *Measures of central tendency and measures of dispersion are important tools.*
- For example: If you stood outside of a movie theater, asked 50 members of the audience if they liked the film they saw, then put your findings on a pie chart, that would be descriptive statistics. In this example, descriptive statistics measure the number of yes and no answers and shows how many people in this specific theater liked or disliked the movie.

Inferential vs Descriptive statistics

Inferential Statistics	Descriptive Statistics
Inferential statistics are used to make conclusions about the population by using analytical tools on the sample data.	Descriptive statistics are used to quantify the characteristics of the data.
Hypothesis testing and regression analysis are the analytical tools used.	Measures of central tendency and measures of dispersion are the important tools used.
It is used to make inferences about an unknown population	It is used to describe the characteristics of a known sample or population.
Measures of inferential statistics are t-test, z test, linear regression, etc.	Measures of descriptive statistics are variance, range, mean, median, etc.

1. Representing Uncertainties in Data: Probability Distributions

What is Uncertainty?

- One of the common features of existing information for machine learning is the **uncertainty** associated with it.
- Real-world data tend to remain incomplete, noisy, and inconsistent. Noise, missing values, and inconsistencies add to the inaccuracy of data.
- Most popular uncertainty management paradigms are based on **probability theory.**
- Probability can be viewed as a numerical measure of the likelihood of occurrence of an outcome relative to the set of other alternatives.
- The set of all possible outcomes is the ***sample space*** and each of the individual outcomes is a ***sample point***.

Probability

"It is a measure of the chance of occurrence of a phenomenon".

We will now discuss some terms which are very important in probability:

- **Random Experiment:** It is an experiment in which all the possible outcomes of the experiments are already known..
- **Sample space:** Sample space of a random experiment is the collection or set of all the possible outcomes of a random experiment.
- **Event:** A subset of sample space is called an event.
- **Trial:** In Trail we have two types of possible outcomes: success or failure with varying Success probability.
- **Random Variable:** A random variable is of two types: Discrete and Continuous variable. In a mathematical way, we can say that a real-valued function X: S -> R is called a random variable where S is probability space and R is a set of real numbers

Probability distributions

- Probability distributions are a function, table, or equation that shows the relationship between the outcome of an event and its frequency of occurrence.
- Probability distributions are helpful because they can be used as a graphical representation of our measurement functions and how they behave.
- When you know how your measurement function have performed in the past, you can more appropriately analyze it and predict future outcomes.

Probability distributions (cont..)

- Probability Mass function
- Probability Density function
- Class conditional probability

Probability Mass function

- Probability mass function can be defined as the probability that a discrete random variable will be exactly equal to some particular value.
- The probability mass function is only used for discrete random variables. For continuous random variables, the probability density function is used which is analogous to the probability mass function.
- The probability mass function is also known as a frequency function. It can be represented numerically as a table, in graphical form, or analytically as a formula

Probability Mass Function Example

Suppose a fair coin is tossed twice and random variable X is number of heads.

1. Tabulate the PMF

2. What is the probability of getting two heads?

Sol: Sample space (S) = [HH, HT, TH, TT]

Let X be the random variable that shows how many heads are obtained. X can take on the values 0, 1, 2 {X=[0,1,2] -> X is discrete random variable}

1.

X	0	1	2
P(X)	1/4	2/4	1/4

→ TABULAR FORMAT

2. P(X=2) -> 1/4

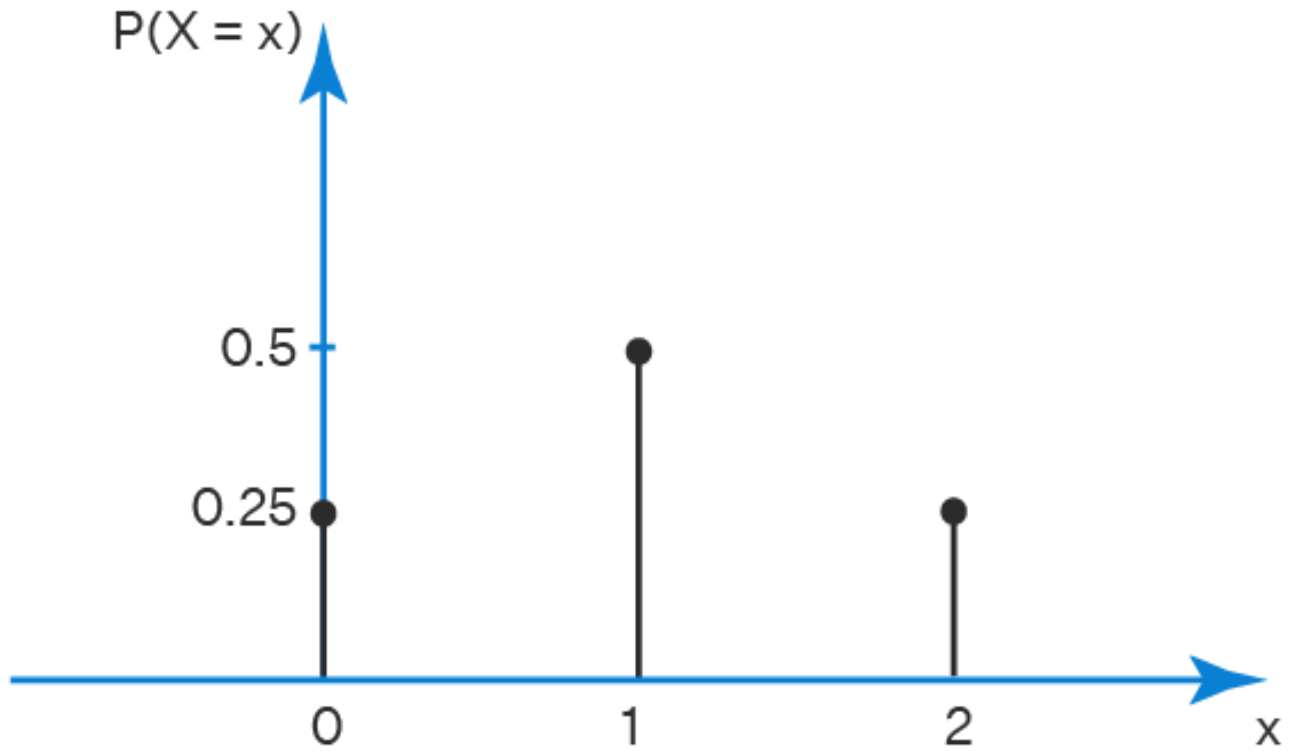
Probability Mass Function Formula

- If X is a discrete random variable with distinct values x1,x2,…xn with their respective probabilities P1,P2,P3……Pn. Then Pi = P(X=xi)= P(xi) is known as probability mass function (p.m.f) if it satisifies the following properties:
  - P(xi) >= 0 , for all i=1,2,3,…..,n
  - $\sum_{i=1}^{n} P(xi) = 1$
  - Here the set of ordered pairs { (xi,P(xi)) / i=1,2,3..n}

    or

    { (x1,P(x1)), (x2,P(x2)), (x3,P(x3)),……., (xn,P(xn))} is called the probability distribution of the random variable X.

Probability Mass Function Graph



Probability Density Function

- Probability density function defines the density of the probability that a continuous random variable will lie within a particular range of values.
- It helps us to understand how the data is distributed across the range and the different outcomes occurring.
- To determine this probability, we integrate the probability density function between two specified points.

Example on PDF

- Consider an example of the time it takes for a customer to complete their purchase at a supermarket checkout counter.

Suppose we observe a group of customers and measure the time each customer spends at the checkout counter. We collect the data and create a histogram, with time intervals on the x-axis (e.g., 1-2 minutes, 2-3 minutes, etc.) and the number of customers falling within each interval on the y-axis.

Now, let's use a probability density function (PDF) to gain more insights into the data.

Assuming the PDF follows an exponential distribution, it will tell us the likelihood of a customer spending a specific amount of time at the checkout counter within a given range.

For instance, let's focus on the time range from 2 to 3 minutes. The PDF will give us information on how probable it is for a customer to spend between 2 and 3 minutes at the checkout counter.

If the PDF indicates a higher probability for times between 2 and 3 minutes, it means that many customers are taking that amount of time to complete their purchases. On the other hand, if the PDF shows a lower probability for this time range, it suggests that fewer customers are taking that specific time duration.

It is important to remember that the PDF gives us relative probabilities for ranges of time intervals and doesn't provide the exact probability of any single customer taking an exact time, such as 2.5 minutes

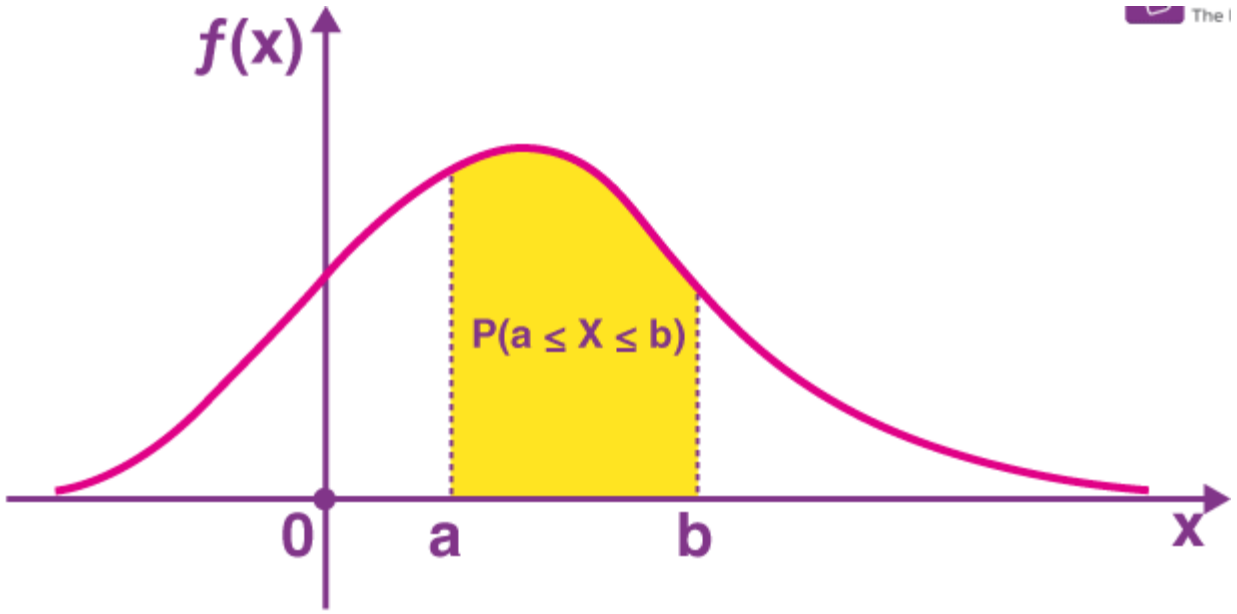
Probability density function formula and graph

- Let X be a continuous random variable. The PDF of X is denoted by f(x) and it satisfies the following properties:
  - f(x)>=0 , for all x, $\mathbf{-\alpha < x < \alpha}$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$ and

$$\mathbf{P(a < X < b) = \int_a^b f(x)\,dx}$$

Or

$$\mathbf{P(a \leq X \leq b) = \int_a^b f(x)\,dx}$$



Conditional probability

- Conditional probability is a probability of occurring an event when another event has already happened.
- Lets suppose, we want to calculate the event A when event B has already occurred, “the probability of A under the conditions of B”, it can be written as:

$$\Rightarrow \mathbf{P(A|B) = P(A \cap B)/P(B)}$$

$$\text{Therefore, } P(A \cap B) = P(B)\, P(A|B) \text{ if } P(B) \neq 0$$

$$= P(A)\, P(B|A) \text{ if } P(A) \neq 0$$

- If the probability of A is given and we need to find probability of B, then it will be given as:

$$\mathbf{P(B|A) = P(B \cap A)/P(A)}$$

Descriptive Measures of Probability Distribution

- Descriptive statistics are broken down into two categories. Measures of central tendency and measures of variability (spread).

(i) **Measure of Central Tendency:**

  - Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way “central” to the set.
  - Mean
  - Median
  - Mode

- **Mean / Average**: Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out

$$x = \frac{12+24+41+51+67+67+85+99}{8} = 55.75$$

- **Median**:Median will be a middle term, if number of terms is odd. Median will be average of middle 2 terms, if number of terms is even.

$$12+24+41+51+67+67+85+99 = 59$$

- **Mode**: Mode is the term appearing maximum time in data set i.e. term that has highest frequency

$$12, 24, 41, 51, 67, 67, 85, 99 \quad = 67$$

(ii) Measures of variability/dispersion:

- **Standard Deviation (Mean deviation / mean absolute deviation)** = sqrt(variance)
- **Variance**: Variance is a statistical measure that tells us how measured data vary from the average value of the set of data. According to the simple terms, it is a measure of how far a set of data i.e. numbers are spread out from their mean i.e. average value.

$$Var(X) = E[X - E(X)^2]$$

$$OrVar(X) = E(X^2) - E(X)^2$$

Example on measuring data dispersion

- Consider the data values of two attributes

  attribute1 = 44,46,48,45 and 47

  attribute 2= 34,46,59,39 and 52

-> if we calculate mean and median of two attributes, we get 46.

-> “However, the first set of values that is the attribute 1 is more concentrated or clustered around the mean/median values whereas the second set of values of attribute 2 is quite spread out or dispersed”.

-> lets prove the above statement by calculating the variances of each attributes.

Covariance:

- Covariance is a measure of the relationship between two random variables, in statistics.
- The covariance indicates the relation between the two variables and helps to know if the two variables vary together.
- In the covariance formula, the covariance between two random variables X and Y can be denoted as Cov(X, Y).
- The variance can be any positive or negative values

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Where,

- X_i is the values of the X-variable
- Y_i is the values of the Y-variable
- $\bar{X}$ is the mean of the X-variable
- $\bar{Y}$ is the mean of the Y-variable
- n is the number of data points

Example: Find covariance for following data set
x = {2,5,6,8,9}, y = {4,3,7,5,6}

Solution:

Given data sets x = {2,5,6,8,9}, y = {4,3,7,5,6} and N = 5

Mean(x) = (2 + 5 + 6 + 8 + 9) / 5

= 30 / 5

= 6

Mean(y) = (4 + 3 +7 + 5 + 6) / 5

= 25 / 5

= 5

Sample covariance $Cov(x,y) = \sum(x_i - x) \times (y_i - y) / (N - 1)$

= [(2 - 6)(4 - 5) + (5 - 6)(3 - 5) + (6 - 6)(7 - 5) + (8 - 6)(5 - 5) + (9 - 6)(6 - 5)] / 5 - 1

= 4 + 2 + 0 + 0 + 3 / 4

= 9 / 4

= 2.25

Population covariance $Cov(x,y) = \sum(x_i - x) \times (y_i - y) / (N)$

= [(2 - 6)(4 - 5) + (5 - 6)(3 - 5) + (6 - 6)(7 - 5) + (8 - 6)(5 - 5) + (9 - 6)(6 - 5)] / 5

= 4 + 2 + 0 + 0 + 3 /

= 9 / 5

= 1.8

Answer: The sample covariance is 2.25 and the population covariance is 1.8.

Descriptive Measures from Data Sample

1. **Range:** It is the difference between the smallest and the largest observation in the sample. It is frequently examined along with the minimum and maximum values themselves.
2. **Mean:** It is the arithmetic average value, that is, the sum of all the values divided by the number of values.
3. **Median:** The median value divides the observations into two groups of equal size—one possessing values smaller than the median and another one possessing values bigger than the median
4. **Mode:** The value that occurs most often, is called the *mode* of data sample.
5. **Variance and Standard Deviation:** The difference between a given observation and the arithmatic average of the sample is called its *deviation*. The variance is defined as the arithmatic average of the squares of the deviations. It is a measure of dispersion of data values; measures how closely the values cluster around their arithmatic average value. A low variance means that the values stay near the arithmatic average; a high variance means the opposite.

   Standard deviation is the square root of the variance and is commonly employed in measuring dispersion. It is expressed in units similar to the values themselves while variance is expressed in terms of those units squared.

Descriptive Measures from Data Sample

6. **Covariance matrix:** Arithmatic average of the matrix $(\mathbf{x}^{(i)} - \boldsymbol{\mu})\,(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T$ gives us *sample covariance matrix*; $i = 1, \ldots, N$ are the N observations in the data sample, and $\boldsymbol{\mu}$ is the arithmatic average of the sample. $\mathbf{x}$ and $\boldsymbol{\mu}$ are both n-dimensional vectors.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \tag{3.19}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\,(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \tag{3.20}$$

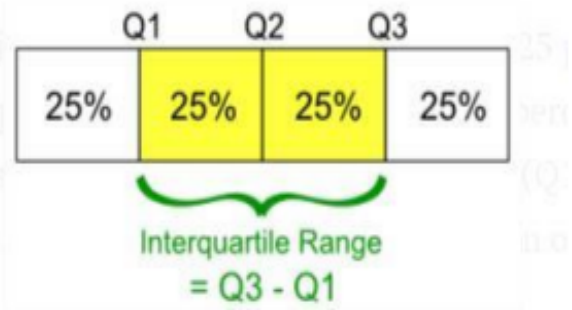
Descriptive Measures from Data Sample

Percentile

Percentile is a way to represent position of a values in data set. To calculate percentile, values in data set should always be in ascending order.

Quartiles

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an **ascending order.**



There are three quartile values. First quartile value is at 25 percentile. Second quartile is 50 percentile and third quartile is 75 percentile. Second quartile (Q2) is median of the whole data. First quartile (Q1) is median of upper half of the data. And Third Quartile (Q3) is median of lower half of the data.

Normal Distributions

- In a normal distribution, data is symmetrically distributed with no skew When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.
- Normal distributions are also called Gaussian distributions or bell curves because of their shape.

Thank you