Classification and Hypothesis Testing

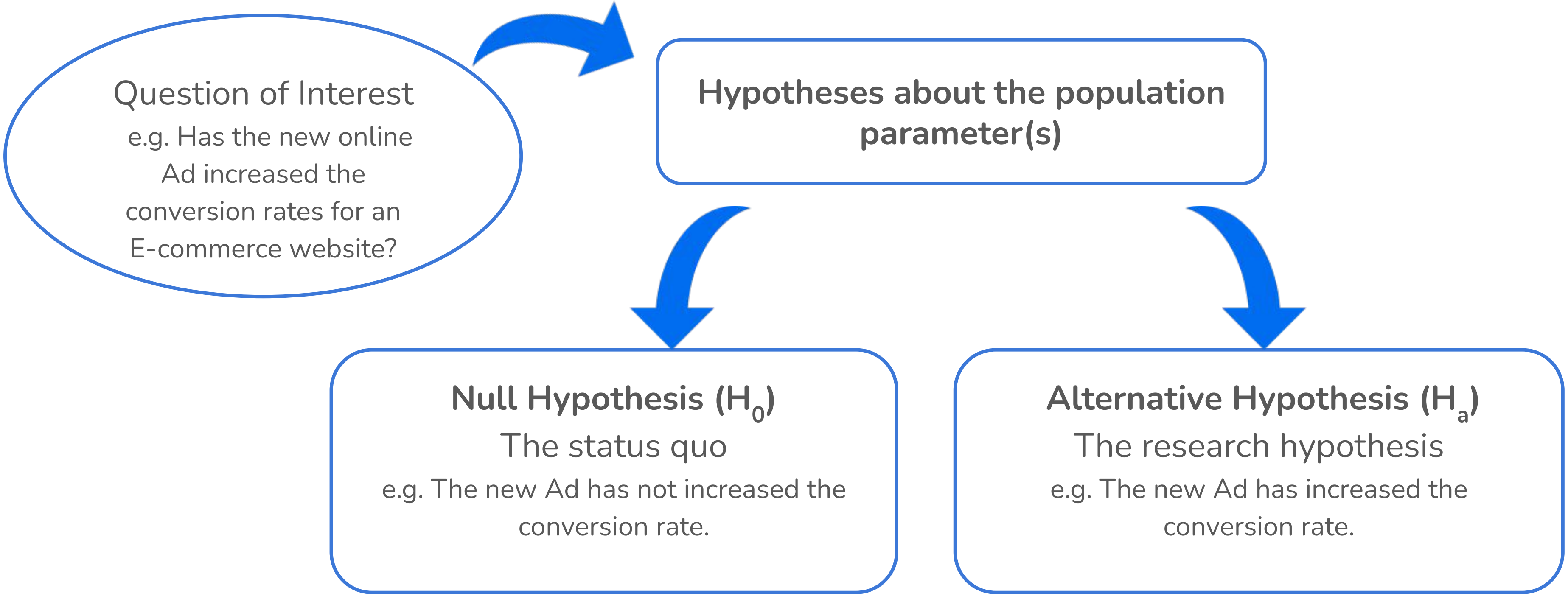
Learning Objective of the Session

1. Hypothesis Testing
    a. Hypothesis Formulation
    b. One Tailed Test vs Two Tailed Test
    c. Type I and Type II Errors
2. Classification
    a. Logistic Regression
    b. SVM and kernel Trick
    c. Perceptron
    d. Decision Tree
    e. Random Forest
    f. Model Evaluation

Discussion Question - Hypothesis Testing

1. What is hypothesis testing and what are different types of hypotheses?
2. What are some of the key terms involved in hypothesis testing?
3. What is the difference between one-tailed and two-tailed tests?
4. What are the steps to perform a hypothesis test?

Introduction to Hypothesis Testing



Question of Interest

e.g. Has the new online Ad increased the conversion rates for an E-commerce website?

Hypotheses about the population parameter(s)

Null Hypothesis (H_0)

The status quo

e.g. The new Ad has not increased the conversion rate.

Alternative Hypothesis (H_a)

The research hypothesis

e.g. The new Ad has increased the conversion rate.

Key terms in Hypothesis Testing

P-Value

- Probability of observing equal or more extreme results than the computed test statistic, under the null hypothesis.
- The smaller the p-value, the stronger the evidence against the null hypothesis.

Level of Significance

- The significance level (denoted by α), is the probability of rejecting the null hypothesis when it is true.
- It is a measure of the strength of the evidence that must be present in the sample data to reject the null hypothesis.

Acceptance or Rejection Region

- The total area under the distribution curve of the test statistic is partitioned into acceptance and rejection region
- Reject the null hypothesis when the test statistic lies in the rejection region, else we fail to reject it

Types of Error

- There are two types of errors - Type I and Type II

Type I and Type II errors

Level of significance = α
Confidence Interval = (1 - α)

	H_0 True	H_0 False
Reject H_0	Type I Error (α)	Correct decision
Fail to reject H_0	Correct decision	Type II Error (β)

Power of the test = (1 - β)

Let’s go through an example

Problem Statement: The store manager believes that the average waiting time for the customers at checkouts has become worse than 15 minutes. Formulate the hypothesis.

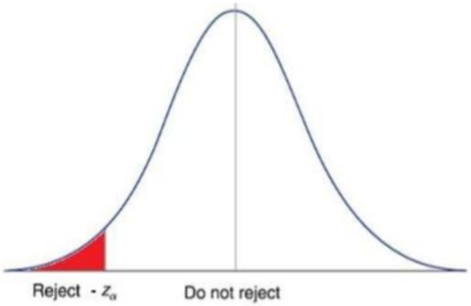
Null Hypothesis (H_0): The average waiting time at checkouts is less than equal to 15 minutes.

Alternate Hypothesis (H_a): The average waiting time at checkouts is more than 15 minutes.

Type I error (false positive): Reject Null hypothesis when it is indeed true. “The fact is that the average waiting time at checkout is less than equal to 15 minutes but the store manager has identified that it is more than 15 minutes”.

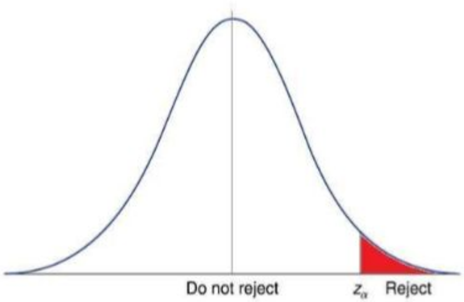
Type II error (false negative): Fail to reject Null hypothesis when it is indeed false. “The fact is that the average waiting time at checkout is more than 15 minutes but the store manager has identified that it is less than equal to 15 minutes”.

One-tailed vs Two-tailed Test



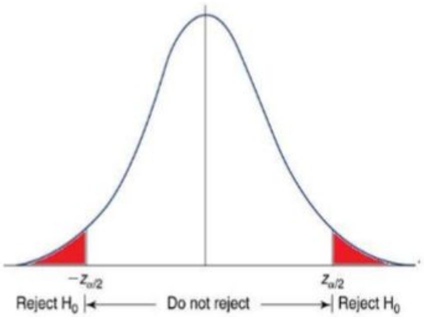
- Lower tail test.
- H_1: μ < ……

Reject H_0 if the value of test statistic is too small



- Upper tail test.
- H_1: μ > ……

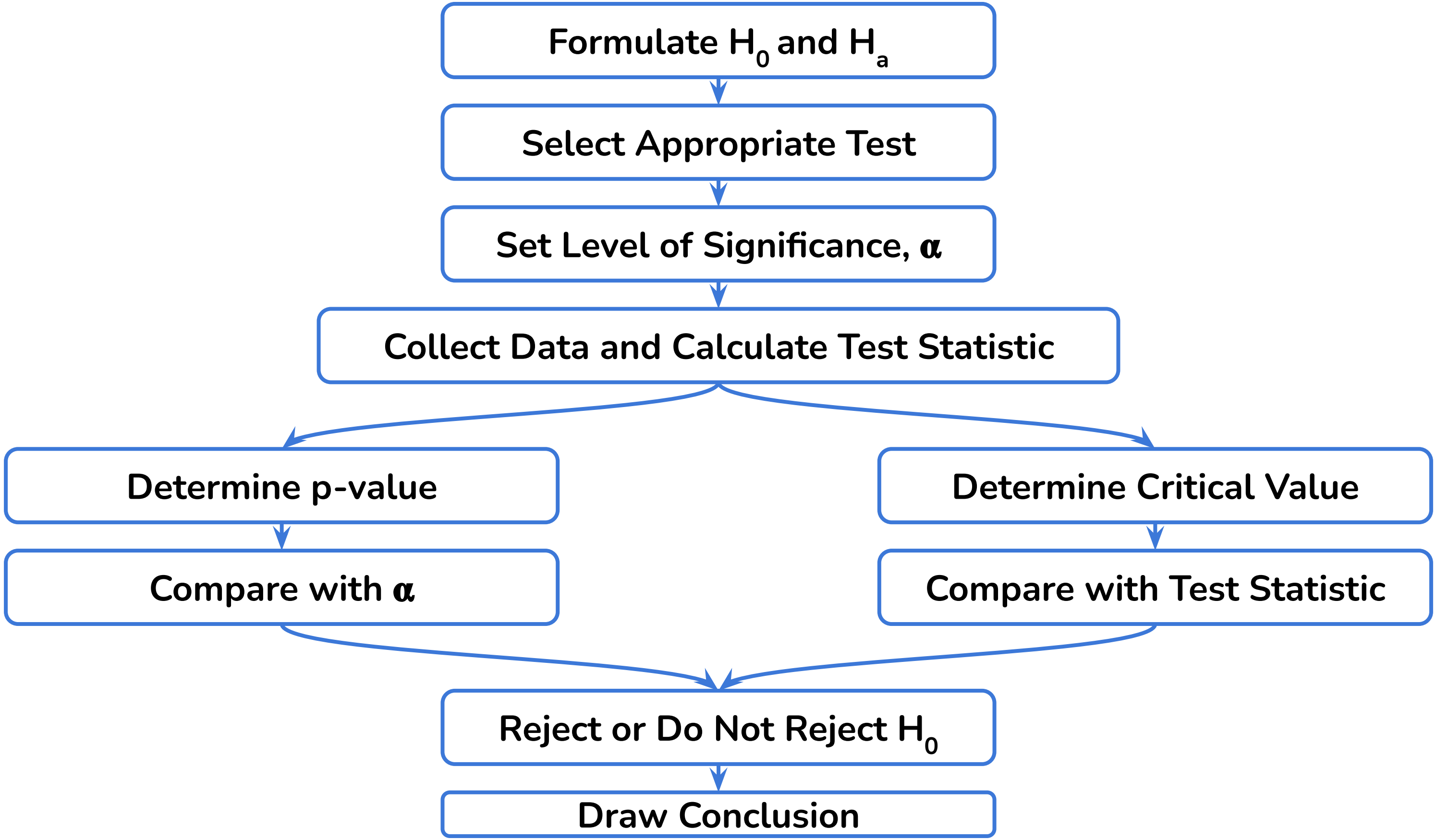
Reject H_0 if the value of test statistic is too large



- Two tail test.
- H_1: μ ≠ ……

Reject H_0 if the value of test statistic is either too small or too large

Hypothesis Testing Steps



Formulate H_0 and H_a

Select Appropriate Test

Set Level of Significance, α

Collect Data and Calculate Test Statistic

Determine p-value

Compare with α

Determine Critical Value

Compare with Test Statistic

Reject or Do Not Reject H_0

Draw Conclusion

Case Study
Hypothesis testing

Why do we use logistic regression?

- Logistic Regression is a supervised learning algorithm which is used for the classification problems i.e. where the dependent variable is categorical
- In logistic regression, we use the sigmoid function to calculate the probability of the dependent variable
- The real life applications of logistic regression are churn prediction, spam detection etc.
- The below image shows how logistic regression is different from linear regression in fitting the model

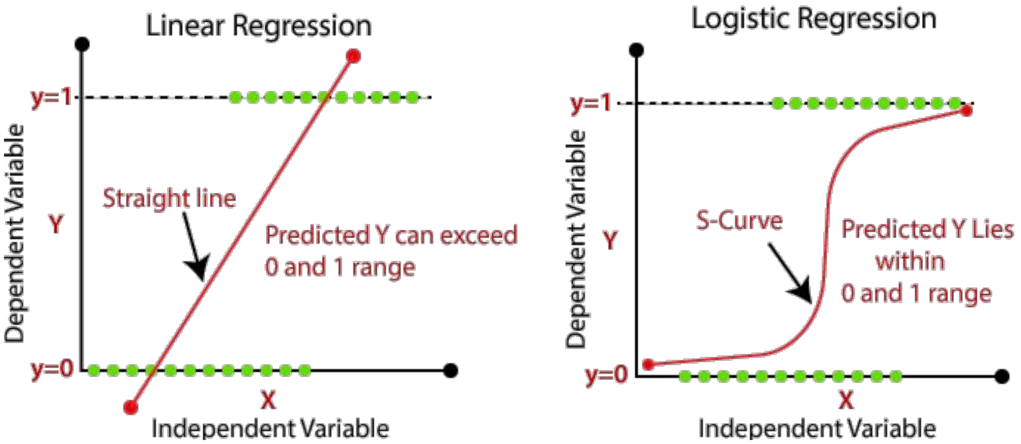
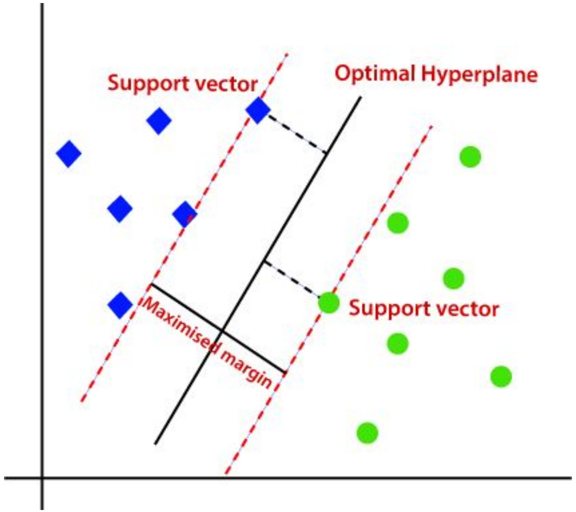


Image Source

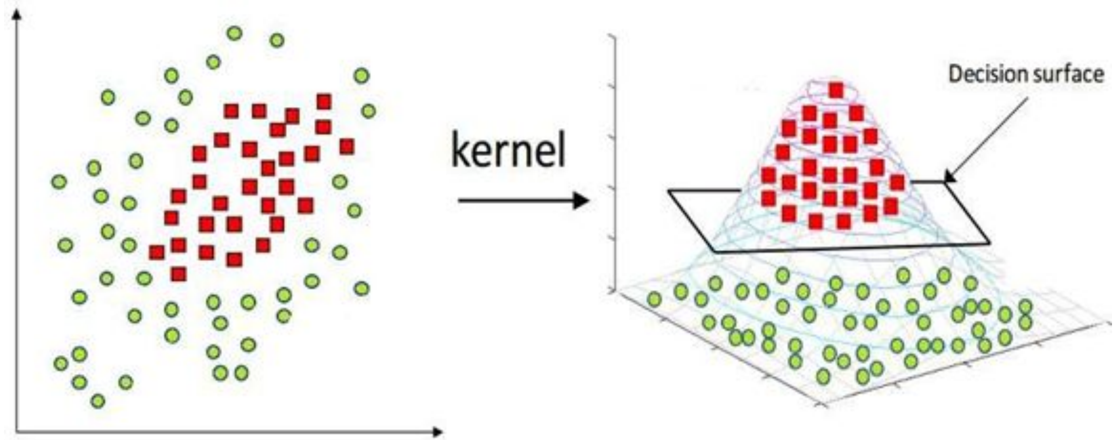
Support Vector Machines

- Support Vector Machines Can be used both for classification and regression tasks, but widely used for classification.
- It Aims at finding a hyperplane in an n-dimensional space that distinctly classifies all the observations in a data
- It is entirely possible to have number of hyperplanes that solve the purpose, we must find the one with maximum margin.
- Dimension of hyperplane is determined by # inputs
    - For 2 input features : line
    - For 3 input features : plane



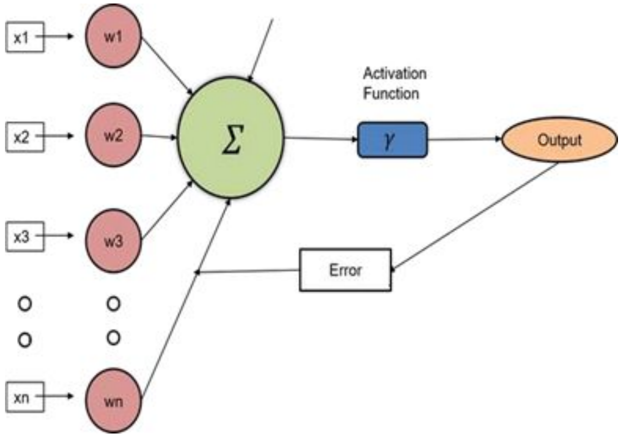
Kernel Trick

- Not all data is good enough to be linearly separable, which makes the job of a SVM difficult
- Not only does it become difficult, but it also leads to high computational costs.
- Kernel trick offers an efficient and less expensive way to transform data
- The trick is to represent the data in a low-dimension space instead of higher dimension.
- This is done by utilizing pairwise comparisons in the original data points.



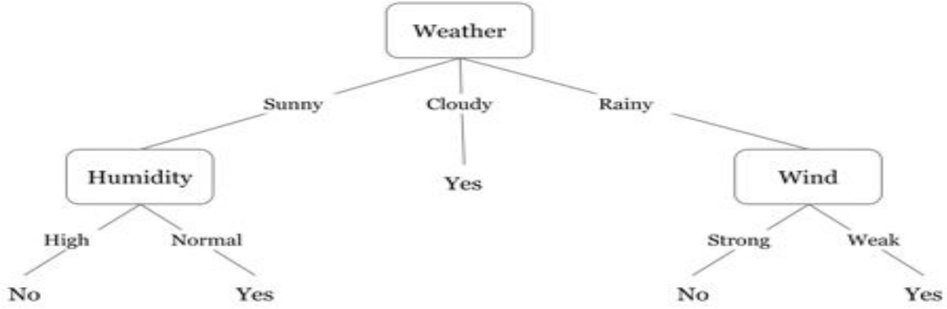
Perceptrons

- Classification technique for binary classification
- It is also called as simplest type of neural network
- Linear classification algorithm i.e., it separates the two classes using a line
- Preferable for data where classes can we well separated using a line
- Coefficients of the model are trained using gradient descent algorithm



Decision Tree

- A decision tree is one of the most popular and effective supervised learning techniques for classification problems, that works equally well with both categorical and continuous variables.
- It is a graphical representation of all the possible solutions to a decision that is based on a certain condition.
- In this algorithm, the training sample points are split into two or more sets based on the split condition over input variables.
- A simple example of a decision tree can be - A person has to decide on going out to play tennis or not by looking at the weather conditions.
  - If it’s cloudy, then the person will go out to play.
  - If it’s sunny, the person will check the humidity level - if normal, the person will go out to play.
  - If it’s rainy, the person further checks the wind speed - if that’s weak, the person will go out to play.



Impurity Measures in Decision Trees

Impurity Measures: Decision trees recursively split features with respect to their target variable's purity. The algorithm is designed to optimize each split such that the purity will be maximized. Impurity can be measured in many ways such as Entropy, Information Gain, etc.

	GINI INDEX	ENTROPY	INFORMATION GAIN
When to use	Classification Tree	Classification Tree	Classification Tree
Formula	$G = 1- \sum^{c}_{i=1}(p_i^2)$	$E = -\sum P(X).\log P(X)$	$IG\ (Y, X) = E(Y) - E(Y\|X)$
Range	0 to 0.5 0 = most pure 0.5 = most impure	0 to 1 0 = most pure 1 = most impure	0 to 1 0 = less gain 1 = more gain
Characteristics	Easy to compute Non-additive	Computationally intensive Additive	Computationally intensive

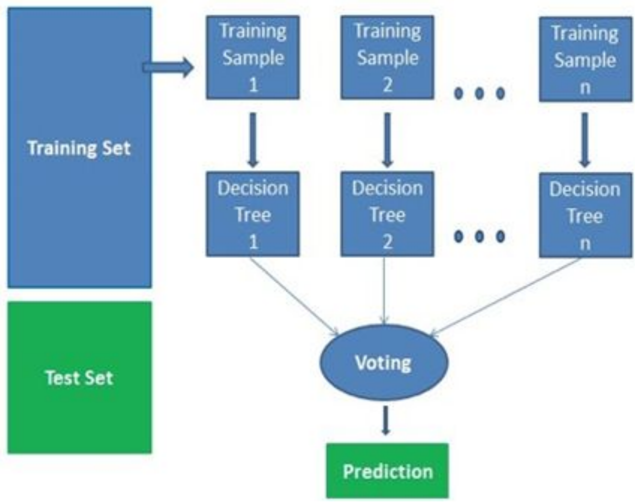
Random Forest

- Random Forest is a supervised machine learning algorithm which can be used for both classification and regression.
- It generates small decision trees using random subsamples of the dataset where the collection of the generated decision tree is defined as forest. Every individual tree is created using an attribute selection indicator such as entropy, information gain, etc.
- In classification, problem voting is done by each tree and the most voted class is considered the final result whereas in case of regression the average method is used to get the final outcome.
- Random Forest is used in various domains such as classification of images, feature selection and recommendation engines.

Steps involved in the Random Forest algorithm

The following steps are involved in this algorithm:

1. Selection of a random subsample of a given dataset.
2. Using attribute selection indicators create a decision tree for each subsample and record the prediction outcome from each model.
3. Applying the voting/averaging method over predicted outcomes of individual models.
4. Considering the final results as the average value or most voted value.



Confusion matrix

It is used to measure the performance of a classification algorithm. It calculates the following metrics:

1. **Accuracy:** Proportion of correctly predicted results among the total number of observations

   Accuracy = (TP+TN)/(TP+FP+FN+TN)

2. **Precision:** Proportion of true positives to all the predicted positives i.e. how valid the predictions are

   Precision = (TP)/(TP+FP)

3. **Recall:** Proportion of true positives to all the actual positives i.e. how complete the predictions are

   Recall = (TP)/(TP+FN)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Case Study
Classification

Happy Learning !