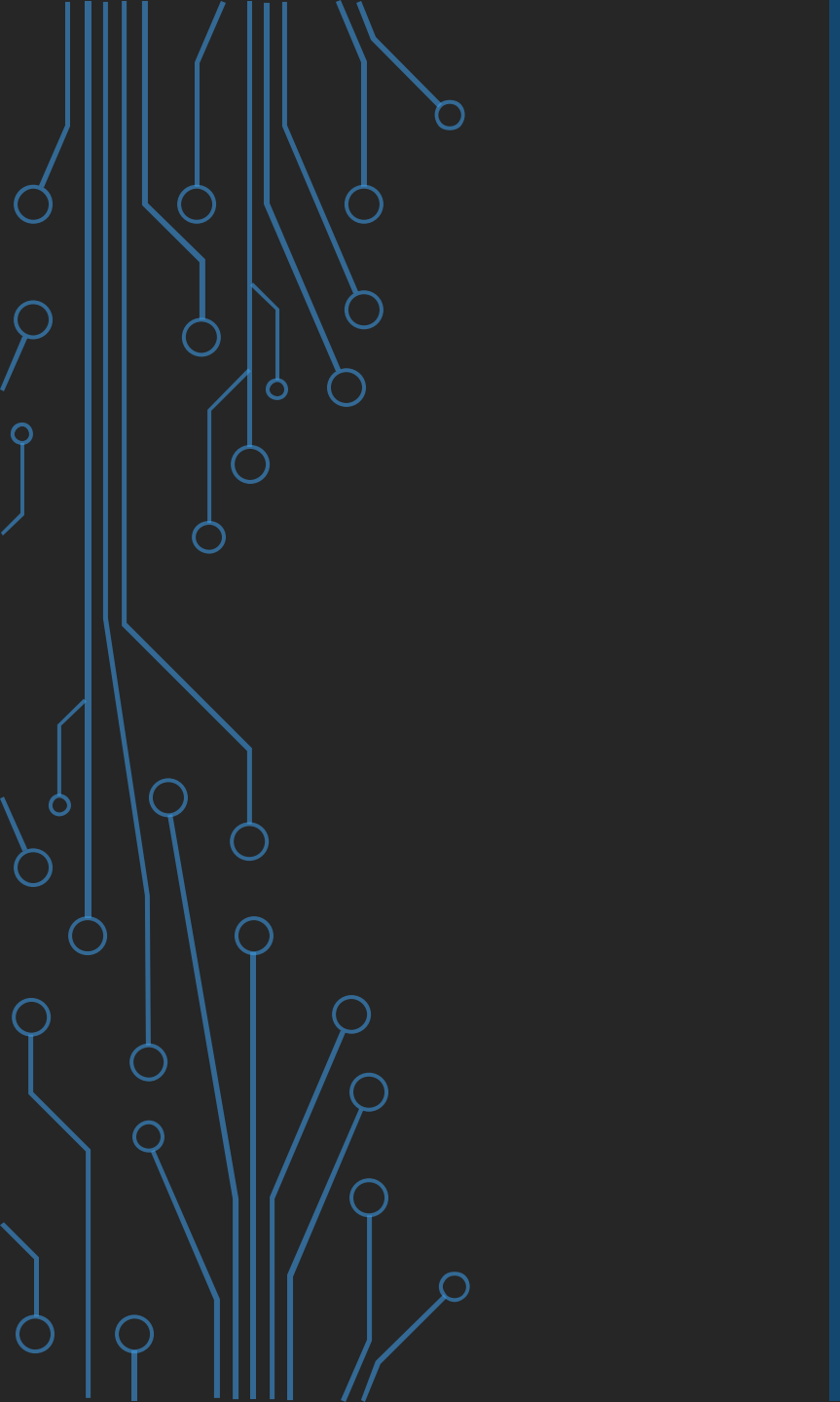
SUPERVISED MACHINE LEARNING (CLASSIFICATION) – PART 2

Machine Learning – Yousef Elbaroudy

GUIDELINES

- Try to focus on the important information mentioned through the session
- Apply what you take on the practical section
- Do not try to memorize everything you got, just learn
- Don’t mind to ask about anything you want to know

Enjoy the Session ☺



PROBABILISTIC LEARNING

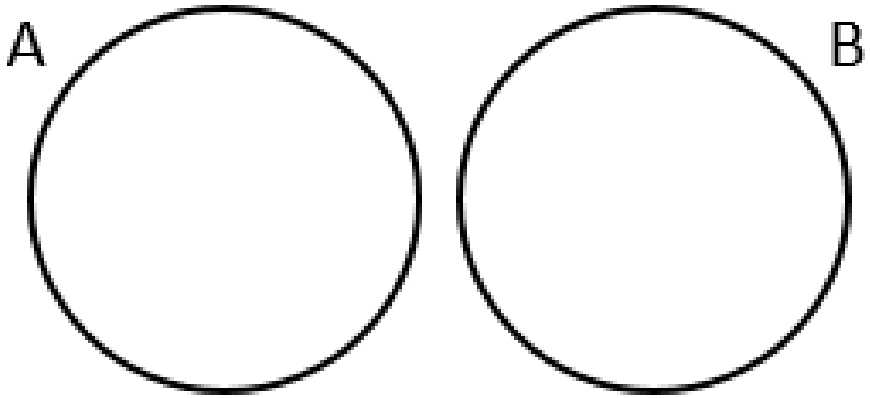
BAYESIAN METHOD

- The technique is used to describe the probability of events, and how probabilities should be revised in the light of additional information.
- These principles formed the foundation for what are now known as **Bayesian methods**.
- It suffices to say that a probability is a number between 0 and 1 (that is, between 0 percent and 100 percent), which captures the chance that an event will occur in the light of the available evidence.

BASIC CONCEPTS OF BAYESIAN METHODS

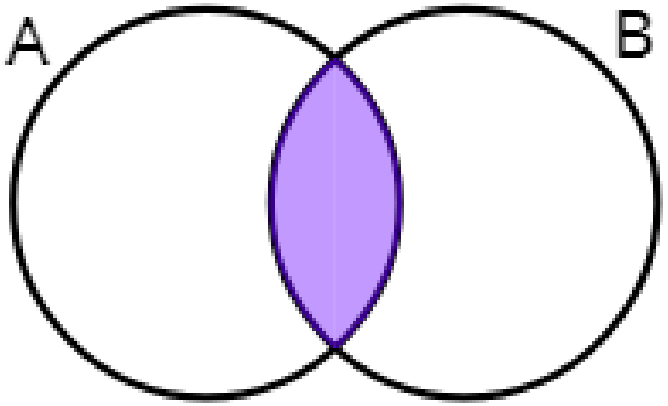
- The probability of an event is estimated from the observed data by dividing the number of trials in which the event occurred by the total number of trials.
- To denote these probabilities, we use notation in the form P(A), which signifies the probability of event A. For example, P(rain) = 0.30 and P(spam) = 0.20.
- The probability of all the possible outcomes of a trial must always sum to 1, because a trial always results in some outcome happening.
- Because an event cannot simultaneously happen and not happen, an event is always **mutually exclusive and exhaustive with its complement**, or the event comprising of the outcomes in which the event of interest does not happen.

Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B)$$

Non-Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

NAÏVE BAYES

- The **Naïve Bayes** theorem is used to predict the probability of a certain class label given a set of features. It's **"naïve"** because it makes a strong assumption that **the features are conditionally independent given the class label**.

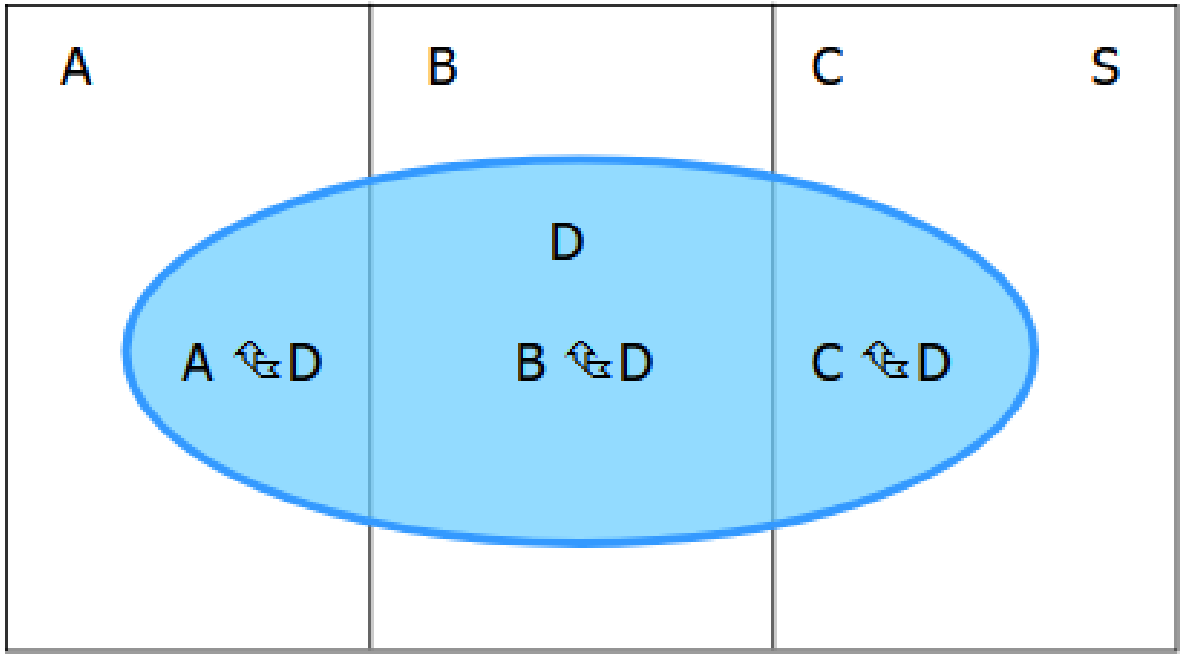
- In other words, it assumes that the presence or absence of a particular feature ***is not influenced by the presence or absence of any other feature***, given the class label.

TYPES OF NAÏVE BAYES

Gaussian Naïve Bayes: the algorithm assumes that the features (attributes) of the data follow a Gaussian (normal) distribution.

Bernoulli Naïve Bayes: each feature is treated as a binary random variable that can take one of two values: 0 (absence) or 1 (presence).

Multinomial Naïve Bayes: refers to the fact that it models the distribution of multiple categories (classes) using a multinomial distribution.



In this Venn Diagram, S is the whole sample space (everything), and D overlaps the other three sets. We will be doing more with this type of Venn.

CONDITIONAL PROBABILITY WITH NAÏVE BAYES

- The relationships between dependent events can be described using **Bayes' theorem**, as shown in the following formula.

$$p(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{p(B|A)p(A)}{P(B)}$$

- The notation $P(A|B)$ is read as the probability of event A, given that event B occurred. This is known as **conditional probability**, since the probability of A is dependent (that is, conditional) on what happened with event B. Bayes' theorem tells us that our estimate of $P(A|B)$ should be based on $P(A \cap B)$, a measure of how often A and B are observed to occur together, and $P(B)$, a measure of how often B is observed to occur in general.

HOW THE MODEL WORKS ?

Training Phase:

- Given a labeled training dataset, the algorithm calculates the probabilities of each feature occurring for each class label.
- It estimates the conditional probabilities of feature counts for each class.

HOW THE MODEL WORKS ? (CONT.)

Prediction Phase:

- When a new document or data point needs to be classified, the algorithm calculates the conditional probabilities of each class label given the observed feature counts.
- It uses Bayes' theorem to calculate these probabilities and predicts the class label with the highest probability.

EXAMPLE

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

likelihood	Outlook w1			Temp w2			Humidity w3		Windy w4		
	Rainy w11	Overcast W12	Sunny w13	Hot w21	Mild w22	Cool w23	High w31	Normal w32	False w41	True w42	total
Play golf (yes)	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	6/9	3/9	9/14
Play golf (no)	3/5	0/5	2/5	2/5	2/5	1/5	4/5	1/5	2/5	3/5	5/14
Total	5/14	4/14	5/14	4/14	6/14	4/14	7/14	7/14	8/14	6/14	14

JOINT PROBABILITY DISTRIBUTION

SO, WEATHER AS THE FOLLOWING (RAINY, COOL, HIGH HUMIDITY AND WINDY) IS IT GOOD OR NOT TO PLAY GOLF ?

$$p(Play\ golf\ (no)|w11 \cap w23 \cap w31 \cap w42) = \frac{p(w11 \cap w23 \cap w31 \cap w42|Play\ golf\ (no))p(Play\ golf\ (no))}{P(w11 \cap w23 \cap w31 \cap w42)}$$

$$p(Play\ golf\ (no)|w11 \cap w23 \cap w31 \cap w42) = (3/5 * 1/5 * 4/5 * 3/5 * 5/14) = 0.020$$

$$p(Play\ golf\ (yes)|w11 \cap w23 \cap w31 \cap w42) = (2/9 * 3/9 * 3/9 * 3/9 * 9/14) = 0.003$$

The probability of not playing golf = 0.020/(0.020+0.003 **= 0.79**

The probability of playing golf = 0.003/ (0.020+0.003)= **0.21**

The Answer is NO

TEST YOUR UNDERSTANDING

Check depend on the following if it is good to play golf or not
(Sunny, mild, Normal Humidity, Not Windy)

likelihood	Outlook w1			Temp w2			Humidity w3		Windy w4		
	Rainy w11	Overcast W12	Sunny w13	Hot w21	Mild w22	Cool w23	High w31	Normal w32	False w41	True w42	total
Play golf (yes)	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	6/9	3/9	9/14
Play golf (no)	3/5	0/5	2/5	2/5	2/5	1/5	4/5	1/5	2/5	3/5	5/14
Total	5/14	4/14	5/14	4/14	6/14	4/14	7/14	7/14	8/14	6/14	14

ADVANTAGES

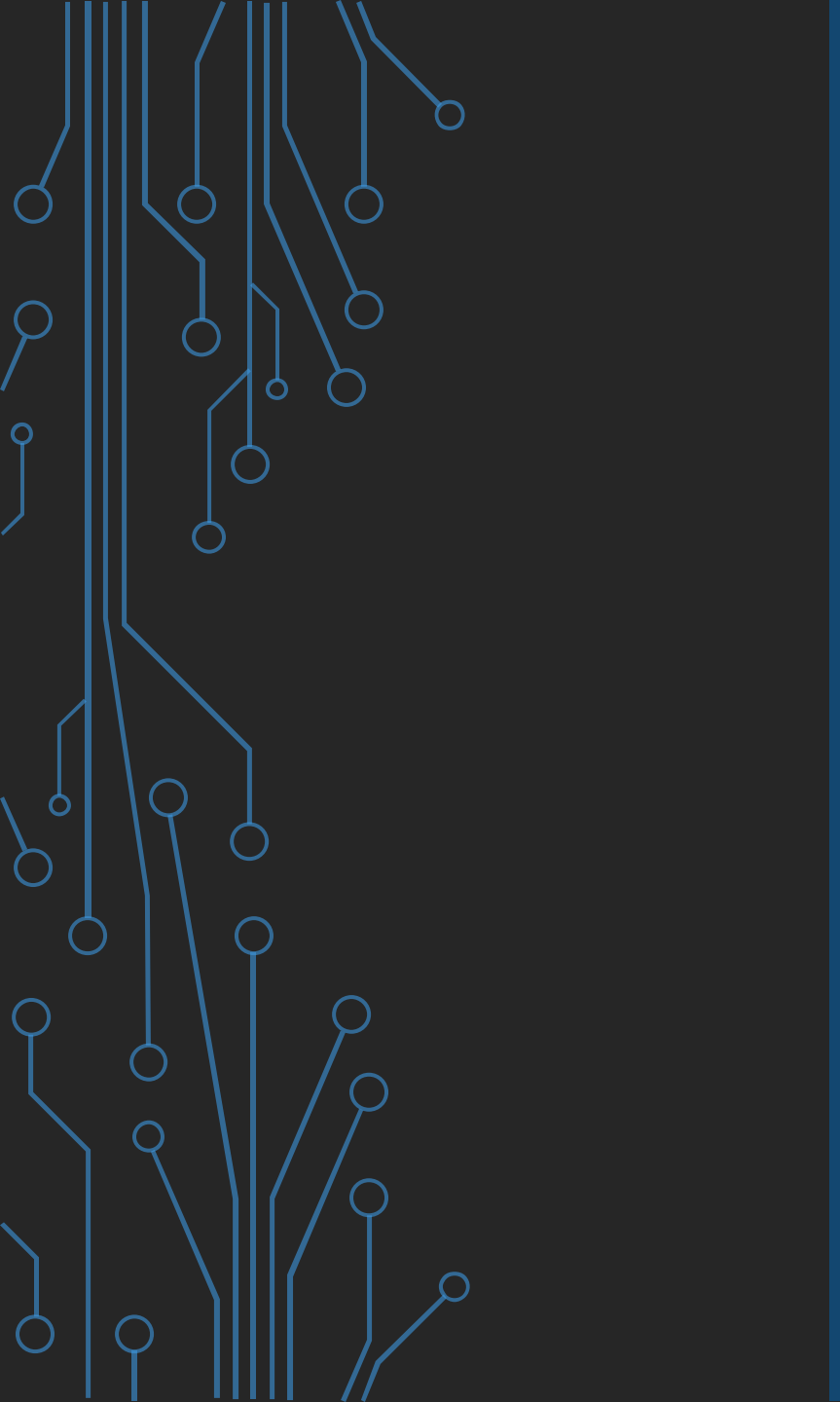
1. **Simplicity and Speed:** Naive Bayes is easy to understand and implement. It's computationally efficient and works well on large datasets. This makes it a good choice for quick prototyping and baseline models.
2. **Scalability:** Naive Bayes can handle high-dimensional data with relatively small sample sizes. It's particularly useful for tasks like text classification where the data is often sparse and high-dimensional.
3. **Real-time Prediction:** Due to its simplicity and efficiency, Naive Bayes can make real-time predictions, making it suitable for applications that require quick decisions.
4. **Strong Performance on Certain Tasks:** Despite its assumptions (naive independence), Naive Bayes can perform surprisingly well on various text classification tasks and situations where the independence assumption holds reasonably well.
5. **Handles Irrelevant Features:** Naive Bayes tends to perform well even when some irrelevant features are present in the data. It is robust to features that may not contribute to the classification decision.
6. **Interpretable Results:** Naive Bayes provides a straightforward way to interpret results and understand the importance of individual features in making predictions.

DISADVANTAGES

1. **Simplistic Assumption:** The key assumption of feature independence can limit the model's ability to capture complex relationships between features. This can lead to suboptimal performance on tasks with strong dependencies between features.
2. **Sensitive to Data Quality:** Naive Bayes can be sensitive to the quality of the training data. Inaccurate or noisy data can negatively impact its performance.
3. **Limited Expressiveness:** While effective for certain tasks, Naive Bayes may not capture subtle patterns and nuances in the data as well as more complex algorithms.
4. **Unsuitable for Continuous Data:** Naive Bayes is designed for discrete data, and its performance may degrade when applied to continuous numerical data without appropriate discretization.
5. **Requires Large Amounts of Data:** Naive Bayes' performance tends to improve with more training data. In cases where data is limited, more complex models might be necessary.
6. **Class Imbalance:** Naive Bayes may struggle with imbalanced datasets, where one class significantly outweighs the other(s). It may disproportionately favor the majority class.

TIME FOR PRACTICALITY (5 MINUTES)

BREAK (10 MINUTES)



SUPPORT-VECTOR MACHINE

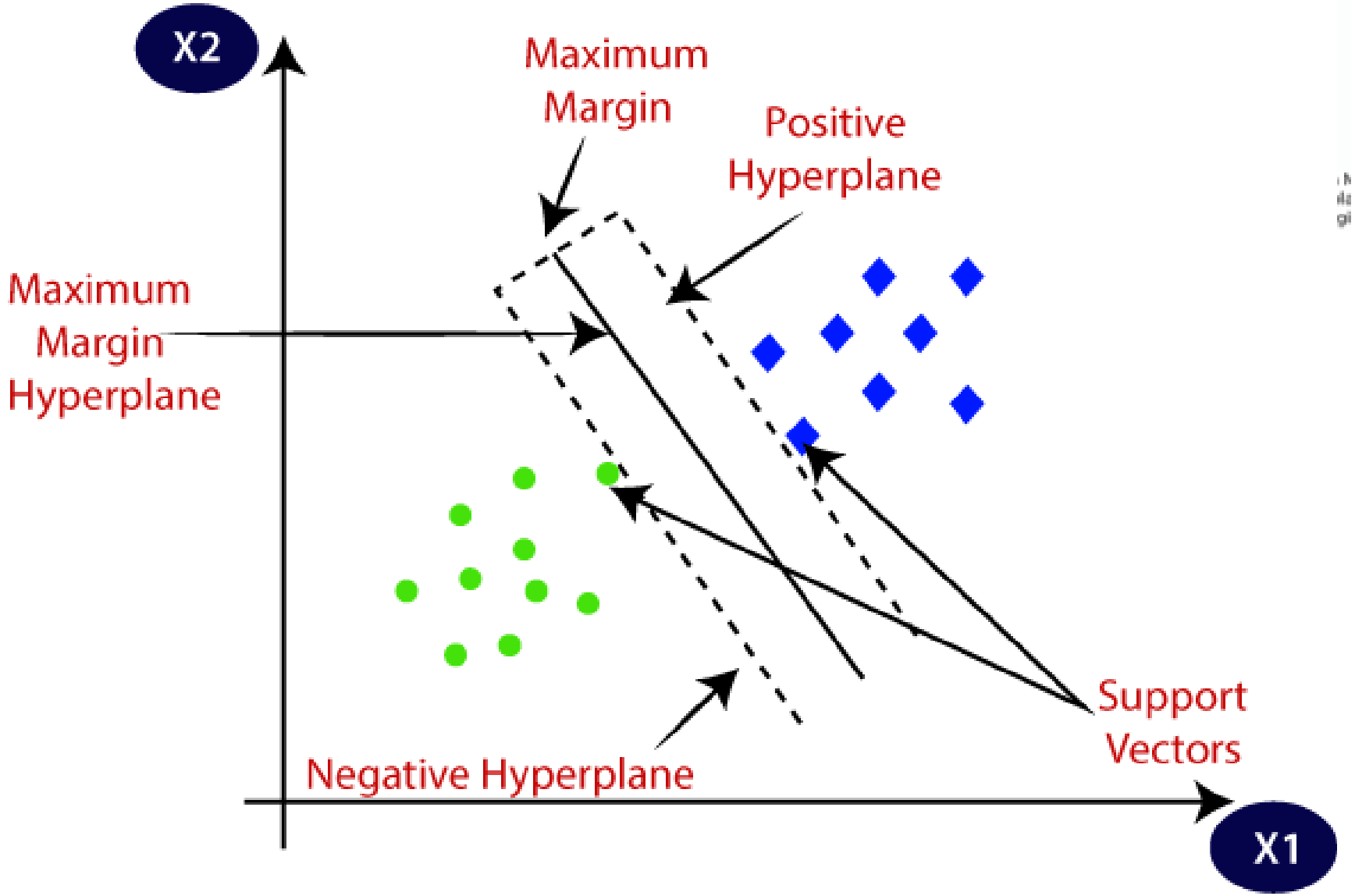
WHAT IS SUPPORT-VECTOR MACHINE ?

- A **Support Vector Machine (SVM)** is a powerful and versatile supervised machine learning algorithm used for both classification and regression tasks.

- SVMs are particularly effective in scenarios where there is a clear separation between classes or when the data is not linearly separable.

What is the main goal ?

- The main goal of an SVM is to find a hyperplane (or decision boundary) that best separates the data into different classes while maximizing the margin between the classes.


X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

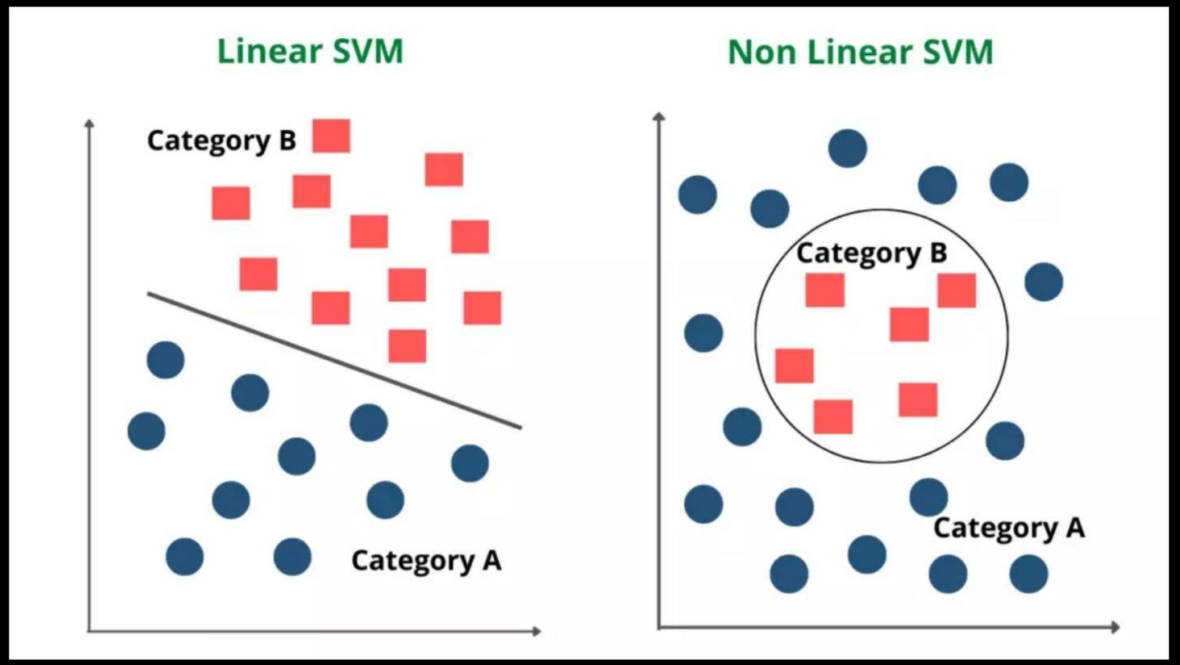
IMPORTANT TERMS

- **Hyperplane:** A hyperplane is a decision boundary that separates data points of different classes. In a binary classification problem, it's a line in two dimensions or a hyperplane in higher dimensions.
- **Support Vectors:** These are the data points that are closest to the hyperplane and have the most influence in determining its position. Support vectors define the margin and play a crucial role in SVM.
- **Margin:** The margin is the distance between the hyperplane and the nearest data points (support vectors) of each class. SVM aims to maximize this margin.
- **Kernel:** A kernel function is used to transform the data into a higher-dimensional space, allowing SVM to find a hyperplane in a transformed feature space. Common kernels include linear, polynomial, radial basis function (RBF), and sigmoid.

TYPES OF SVM

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

SVM in ML


Linear SVM
Category B
Category A
Non Linear SVM
Category B
Category A

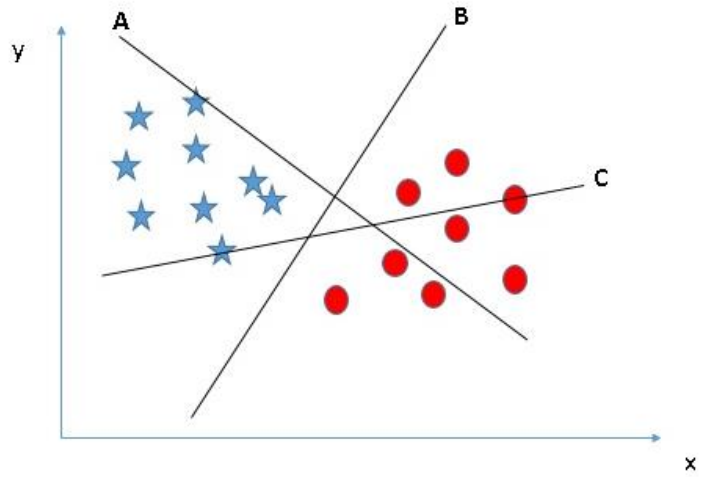
EXISTENCE OF THE KERNELS (HYPERPLANE)

The mathematical function used for the transformation is known as the ***kernel function.*** Following are the popular functions.

- Linear
- Polynomial
- Radial basis function (RBF)
- Sigmoid

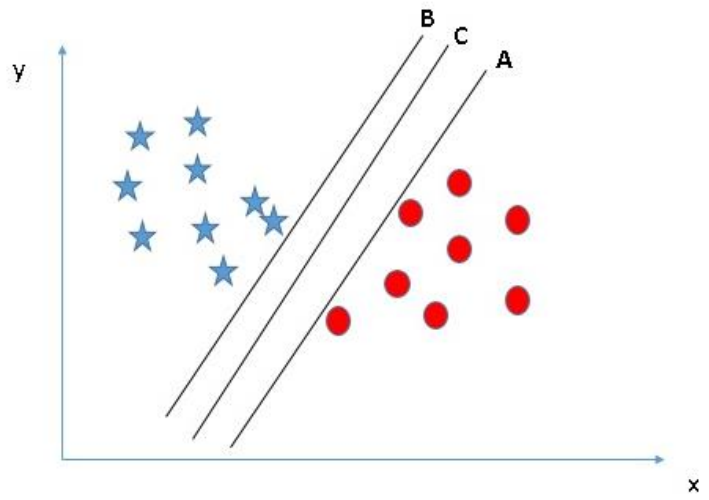
WHY "MARGIN" ?

- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B, and C). Now, identify the right hyper-plane to classify stars and circles.



WHY "MARGIN" ? (CONT.)

- **Identify the right hyper-plane (Scenario-2):** Here, we have three hyper-planes (A, B, and C), and all segregate the classes well. Now, How can we identify the right hyper-plane?

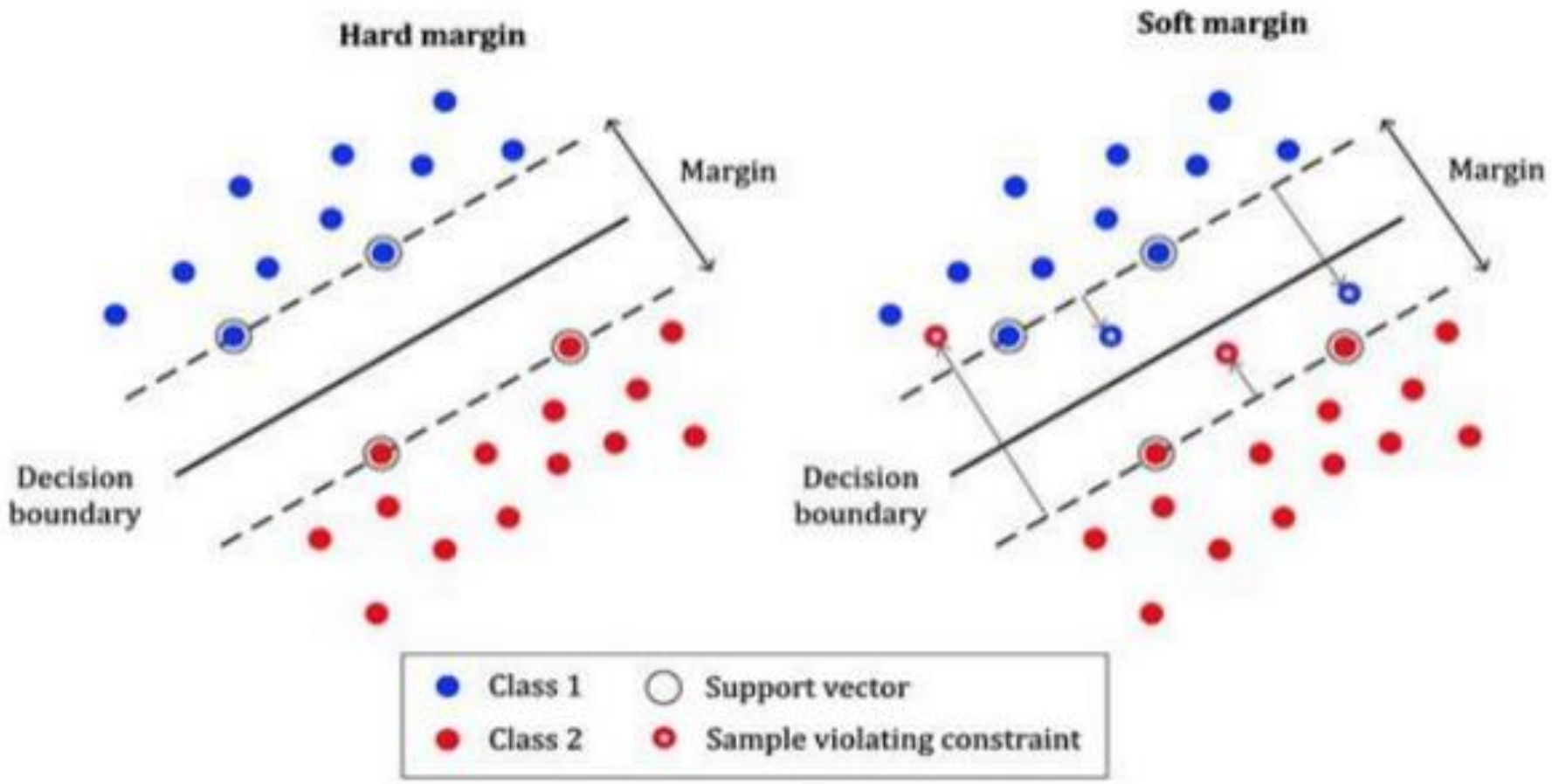


THE LEARNING PROCESS

- The distance of the vectors from the hyperplane is called the margin which is a separation of a line to the closest class points.

- We would like to choose a hyperplane that **maximizes the margin** between classes.

  - Soft Margin
  - Hard Margin



The cost function is to maximize the margin

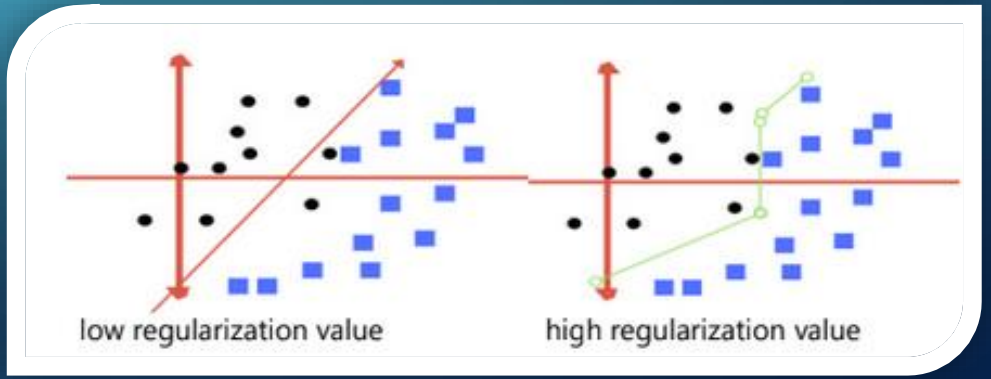
SOFT VS. HARD MARGIN

1- Soft Margin – As most of the real-world data are not fully linearly separable, we will allow some margin violation to occur which is called soft margin classification. It is better to have a large margin, even though some constraints are violated. Margin violation means choosing a hyperplane, which can allow some data points to stay on either the incorrect side of the hyperplane and between the margin and correct side of the hyperplane.

2- Hard Margin – If the training data is linearly separable, we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible.

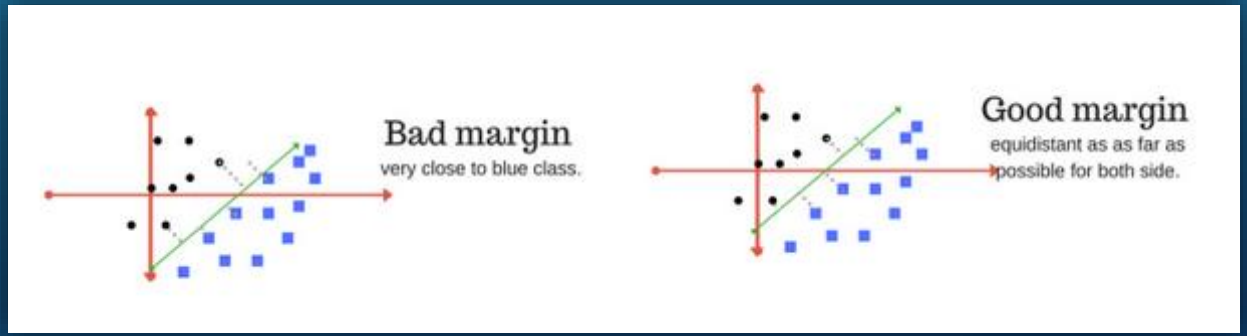
REGULARIZATION

- The Regularization parameter (often termed as C parameter in python‘s sklearn library) tells the SVM optimization how much you want to avoid misclassifying each training example.
- For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.
- Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.



GOOD FITTING

- A good margin is one where this separation is larger for both the classes. Images below gives to visual example of good and bad margin.



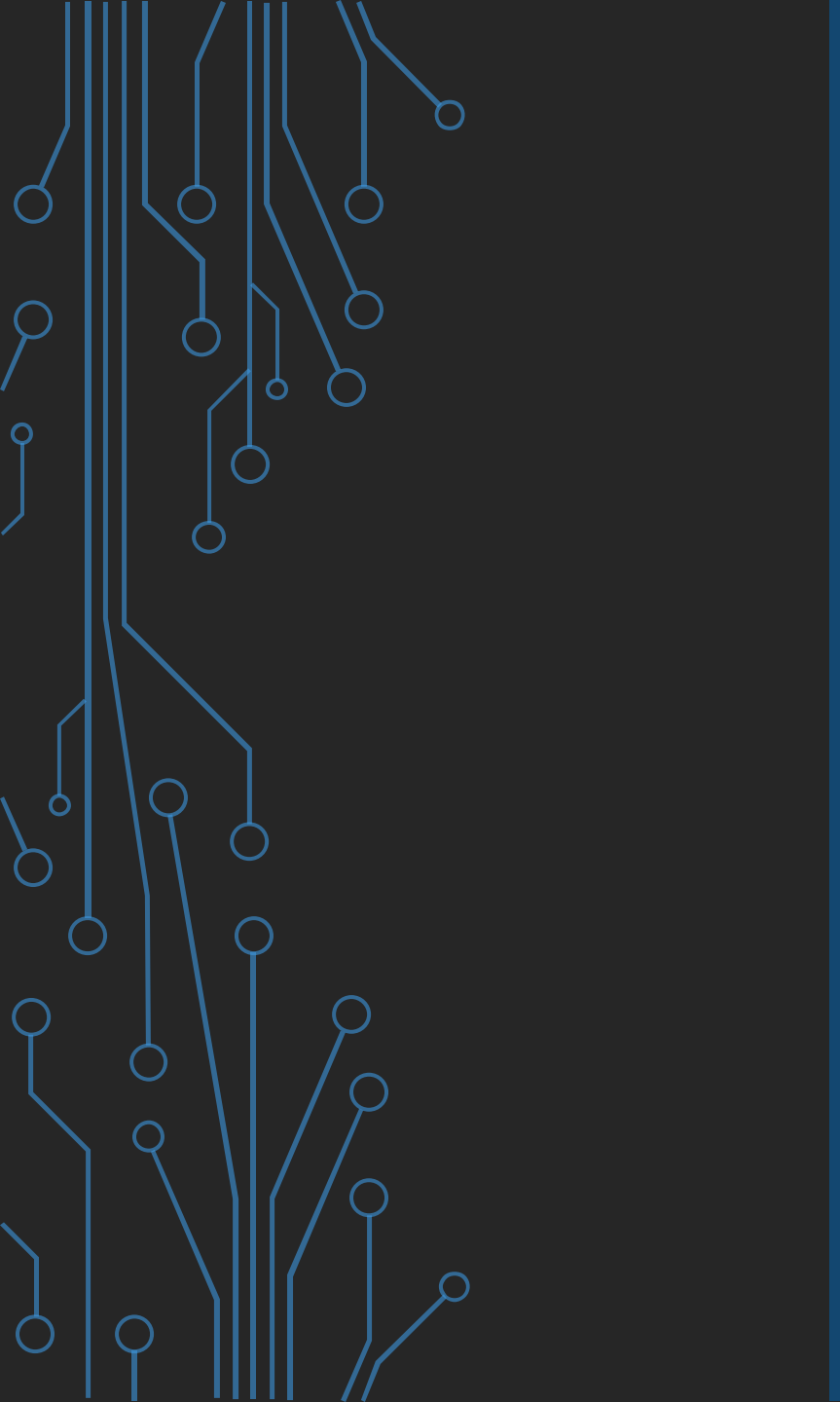
ADVANTAGES

- Effective in high-dimensional spaces and with small to medium-sized datasets.
- Works well in cases where classes are not linearly separable through the use of kernel functions.
- Tends to generalize well and avoid overfitting, especially when the regularization parameter C is properly tuned.

LIMITATIONS

- SVMs can be computationally intensive, especially for large datasets.
- Choosing the appropriate kernel and hyperparameters requires careful tuning and experimentation.
- Interpreting the model's decisions can be challenging, especially in high-dimensional spaces.
- SVMs may not perform well when dealing with noisy data or data with a high degree of overlap between classes.

TIME FOR PRACTICALITY (5 MINUTES)



EVALUATION METRICS

EVALUATION METRICS

Each Machine Learning Problem has it is own Evaluation Metrics

CLASSIFICATION EVALUATION METRICS

- **Evaluation metrics** for classification are used to assess the performance of a classification model by comparing its predictions to the actual class labels in a dataset.

- These metrics help quantify how well the model is performing and provide insights into its strengths and weaknesses.

Accuracy	Precision	Recall	F1-Score	ROC	AUC

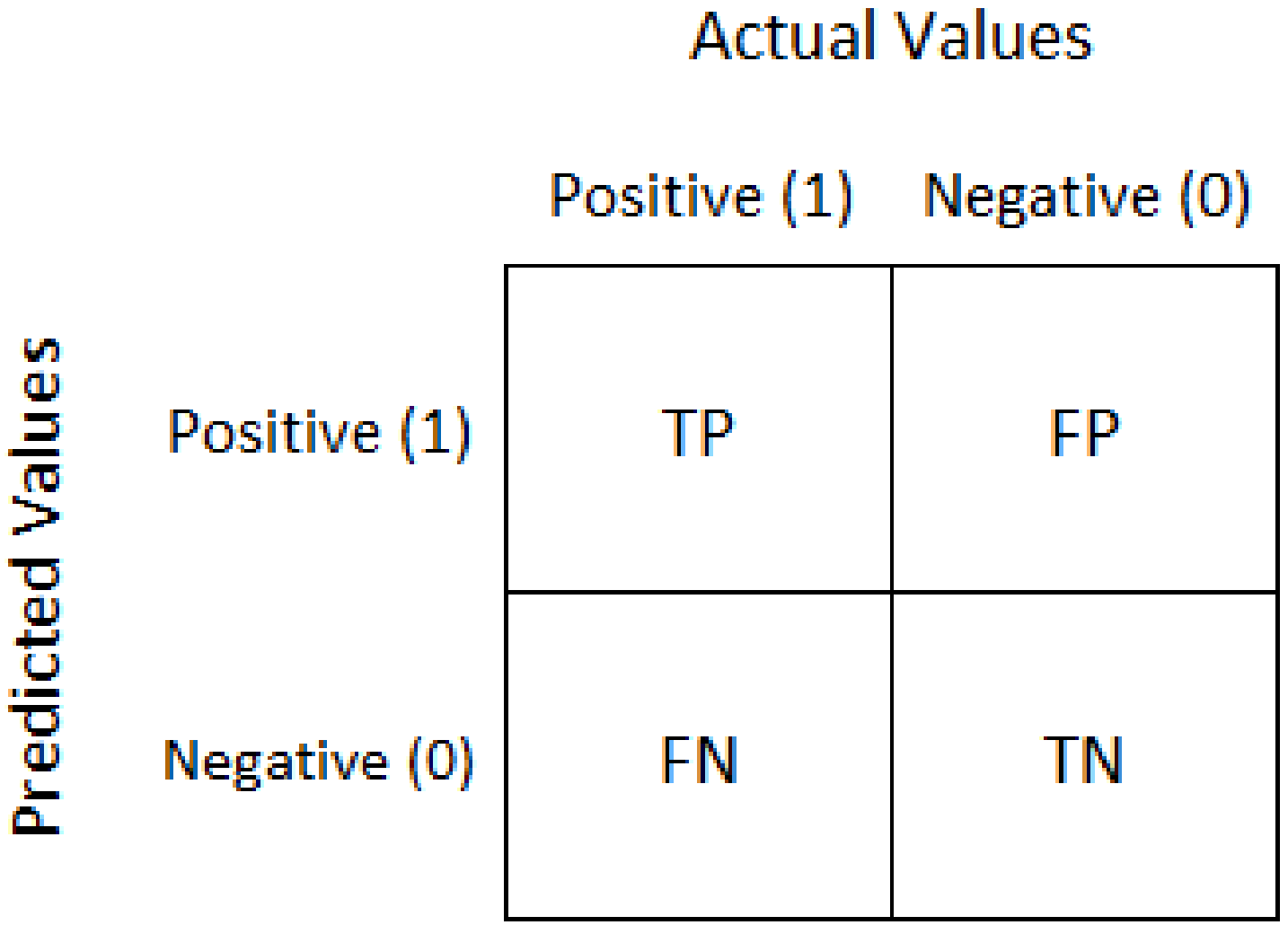
ACCURACY

Accuracy: Accuracy measures the proportion of correctly predicted instances out of the total instances in the dataset. It's a simple and intuitive metric but may not be suitable for imbalanced datasets.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

CONFUSION MATRIX

Confusion Matrix: A confusion matrix provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions. It's a useful tool for understanding the distribution of prediction errors.


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

Precision: Precision measures the proportion of true positive predictions (correctly predicted positives) out of all predicted positives. It focuses on the correctness of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Used for Medical Diagnosis

PRECISION

Recall (Sensitivity or True Positive Rate): Recall measures the proportion of true positive predictions out of all actual positives. It focuses on the completeness of positive predictions.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Used for information Retreiving

RECALL

F1-SCORE

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is especially useful when class imbalance is present.

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

SPECIFICITY

Specificity (True Negative Rate): Specificity measures the proportion of true negative predictions out of all actual negatives. It's the complement of the false positive rate.

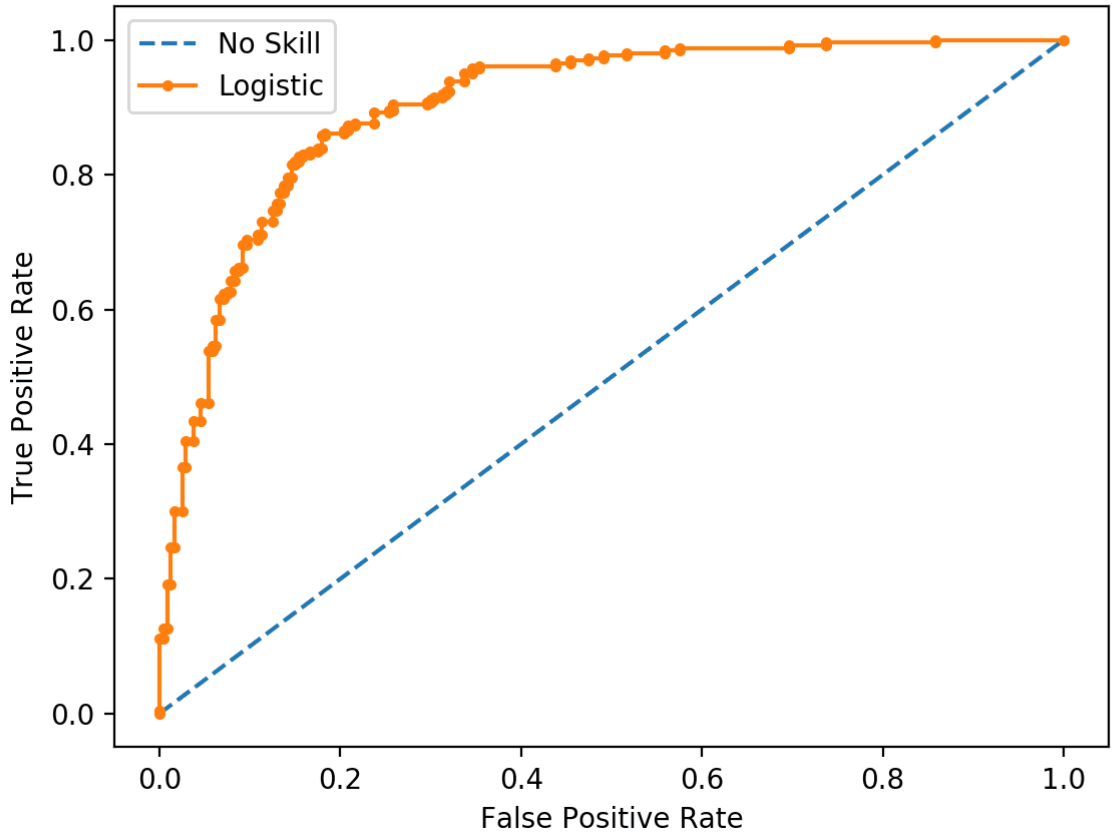
$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

RECEIVER OPERATION CHARACTERISTIC (ROC)

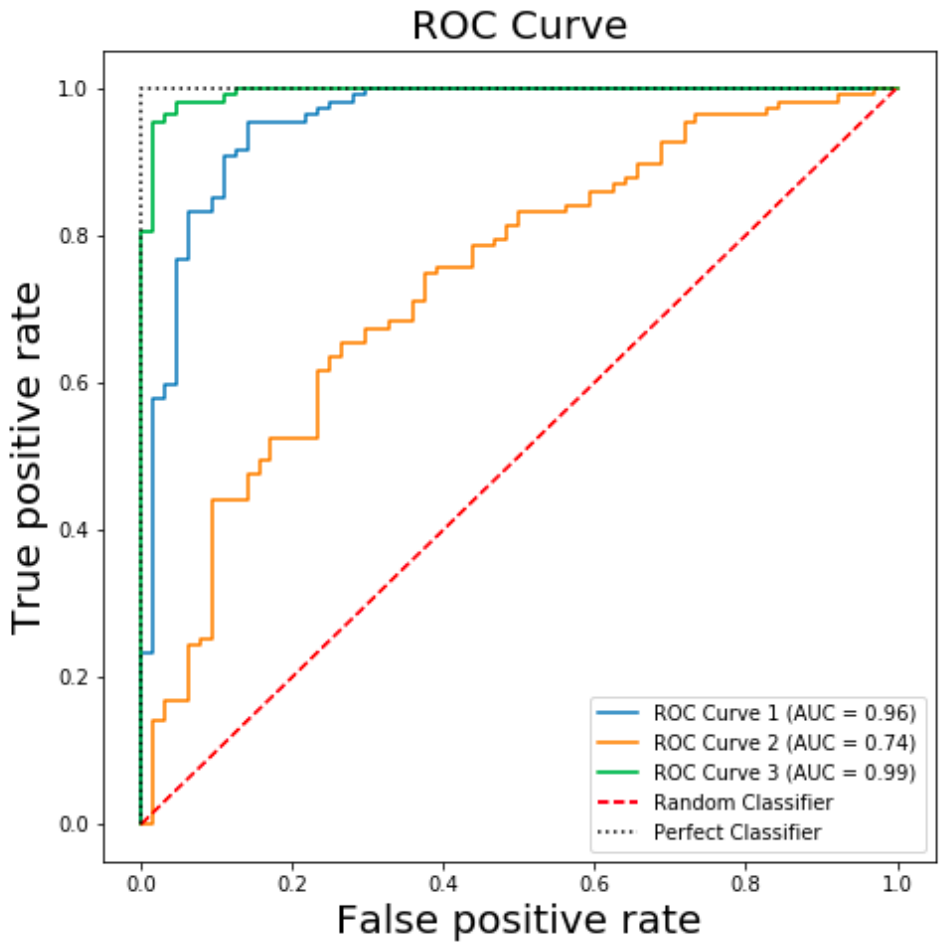
- **ROC curve** is a graphical representation of the true positive rate (recall) versus the false positive rate (1 – Specificity) for different threshold values.

- **The area under the ROC curve (AUC-ROC)** is a common metric that quantifies the overall performance of a classifier.

AUC-ROC CURVE



AUC-ROC CURVE



TIME FOR PRACTICALITY (5 MINUTES)

QUESTIONS


THANK YOU