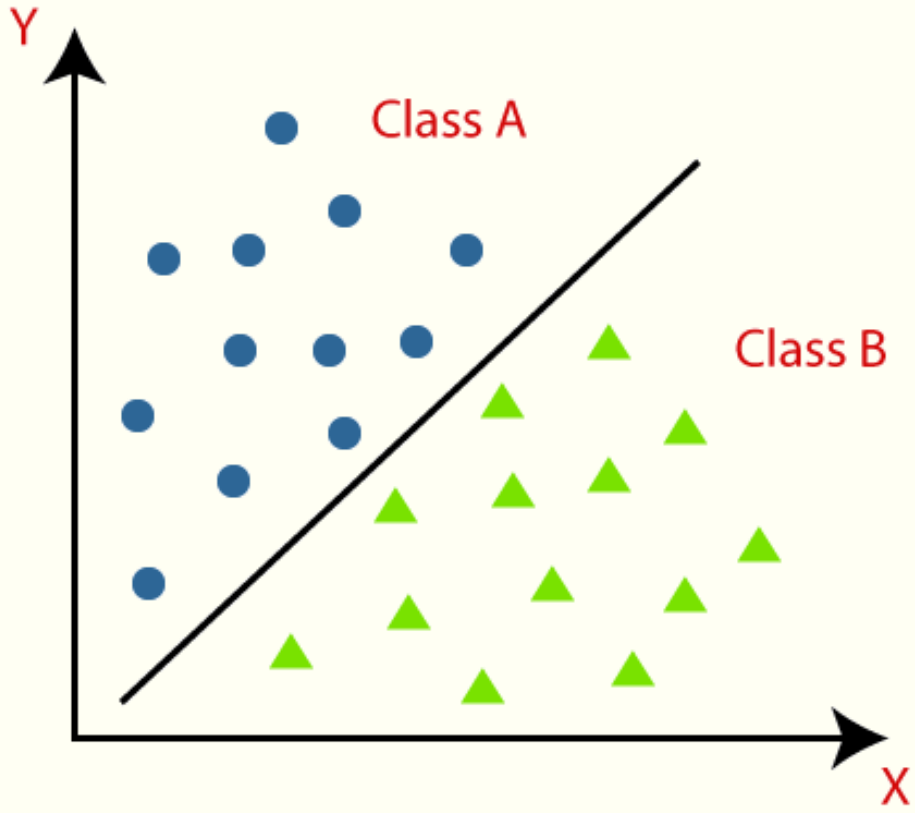
A RECAP OF MACHINE LEARNING CLASSIFIERS

by Siddharth Dixit



MAT-494: Deep Learning
(Monsoon-2020)

Absolute Basics

- Classification is a predictive modeling problem that involves assigning a class label to each observation.

1) **Binary Classification**: All observations belong to one of two classes.

2) **Multiclass Classification:** All observations belong to one of three classes.

For example, we may collect measurements of a flower and classify the species of flower (*label*) from the measurements ~ IRIS DATASET

- We may alternately use predict the probability of class membership (*Soft Classifiers*) instead of a strict class label (*Hard Classifiers*) which allows you to quantify uncertainty in a prediction across a range of options and allow the user to interpret the result in the context of the problem.

Types of Classifications

- **Balanced Classification:-**

When we have approximately equal instances of each class in the dataset.

- **Imbalanced Classification:-**

Distribution of examples across the known classes is biased or skewed in the dataset.

Examples- Credit Card Fraud, Spam Detection, and Churn Prediction

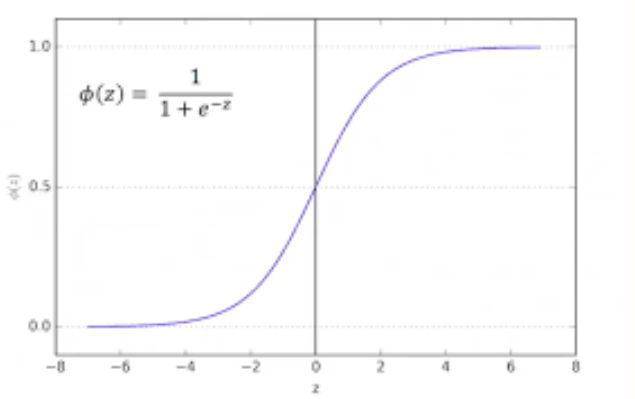
Most Frequently used Models for Balanced Classification

- Logistic Regression (LDA, QDA for MultiClass)
- K- Nearest Neighbours (kNN)
- Naïve Bayes
- Support Vector Machines (with different kernels)
- Tree Based Methods:- Decision Trees, Random Forests, Gradient Boosting (XGBoost)
- Several different types of Neural Networks (which we'll study as this course progresses)

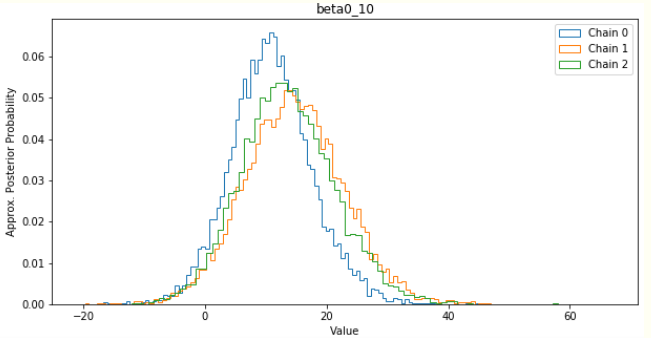
Logistic Regression

- Derives its name from logistic/sigmoid function developed by statisticians to describe properties of population growth in ecology.
- Can only be used for Binary Classification.

Equation:- P(Y=1|X) = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))



- Training is done using Maximum Likelihood Estimation to find the optimal coefficient values (point estimates).
- Better version:- Bayesian Logistic Regression, which gives posterior distributions instead of point estimates for coefficients.

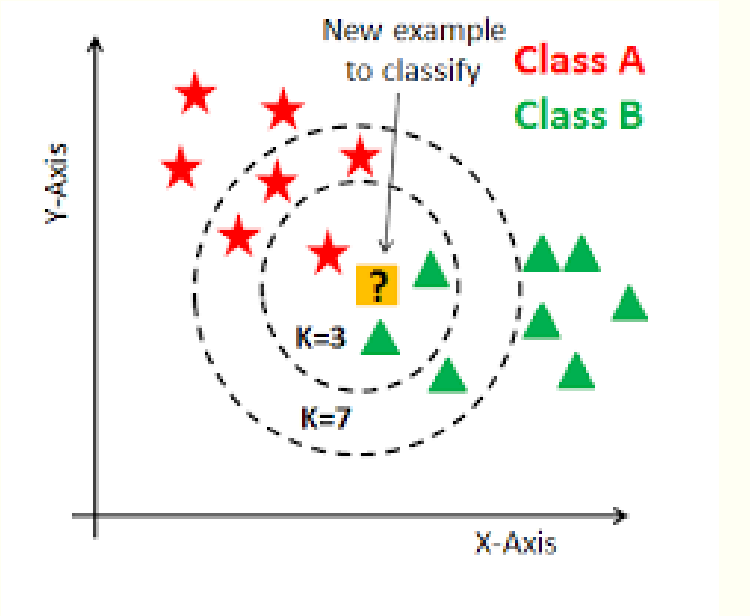


K-Nearest Neighbours

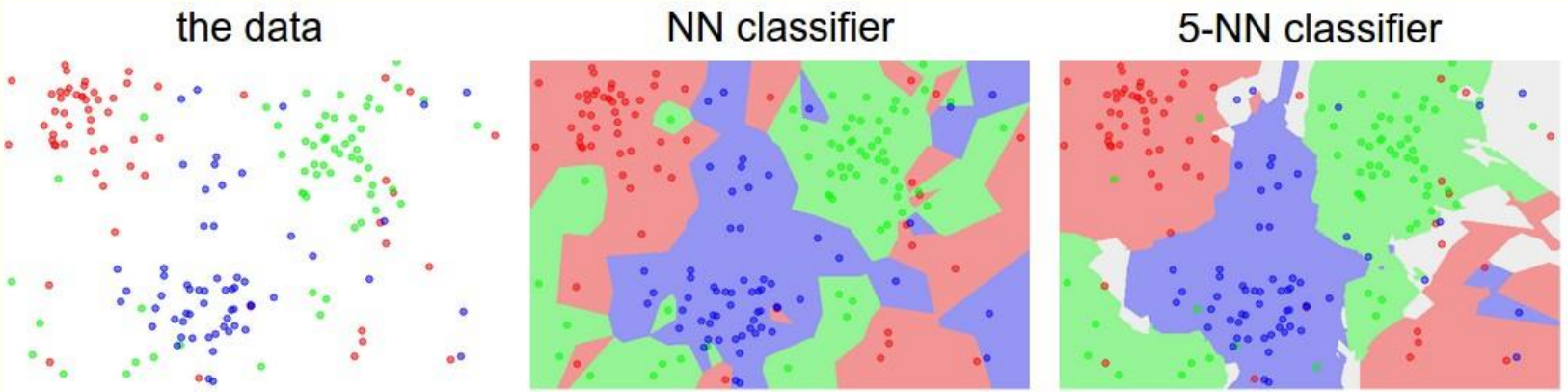
- Parametric or Non-Parametric model?

Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.

*- **Artificial Intelligence: A Modern Approach by Russel & Norvig***



Source: Datacamp



How kNN decision boundaries look like

Source: Stackoverflow

Naïve Bayes Classifier

- Parametric Algorithm only used for classification. Called Naïve (Idiot) because assumes independence amongst features which rarely happens in real data, i.e. that the features do not interact. Still, the algorithm performs surprisingly well on data where this assumption does not hold. Ex- Spam Detection and Document Classification using text data.

Uses Bayes Theorem: $P(Fraud|X) = (P(X|Fraud) * P(Fraud)) / P(X)$

- $P(Class|X)$ is the probability of a particular class given the data X. This is called the **posterior probability**.
- $P(X|Class)$ is the probability of data d given that we observed a particular class.
- $P(Class)$ is the probability of a particular class (regardless of the data). This is called the **prior probability** of class.
- $P(X)$ is a marginal and is the probability of the data (regardless of the class).

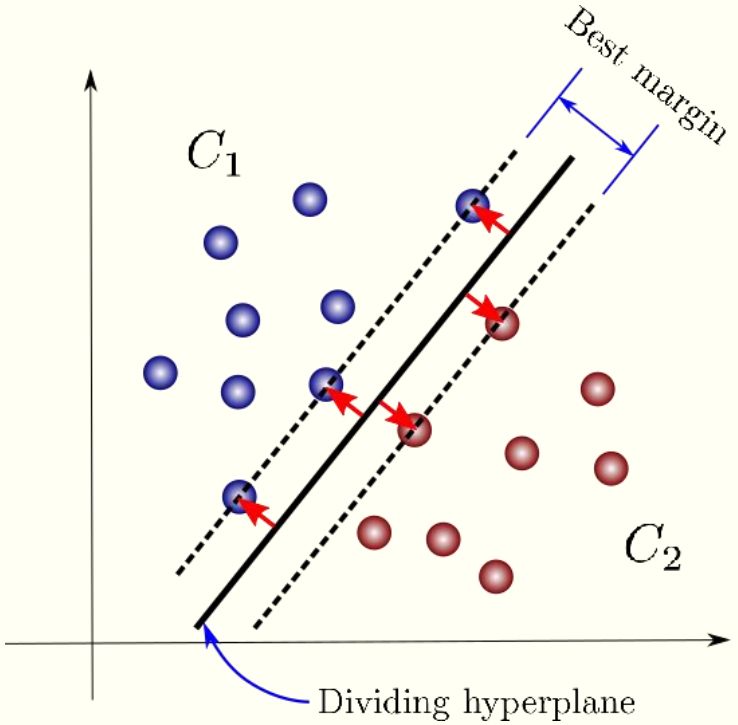
After calculating the posterior probability for several different classes, you can select the class with the highest probability. This is the maximum probable class and may formally be called the Maximum A Posteriori (MAP) class.

$$MAP(Class) = max((P(X|Fraud) * P(Fraud)) / P(X))$$

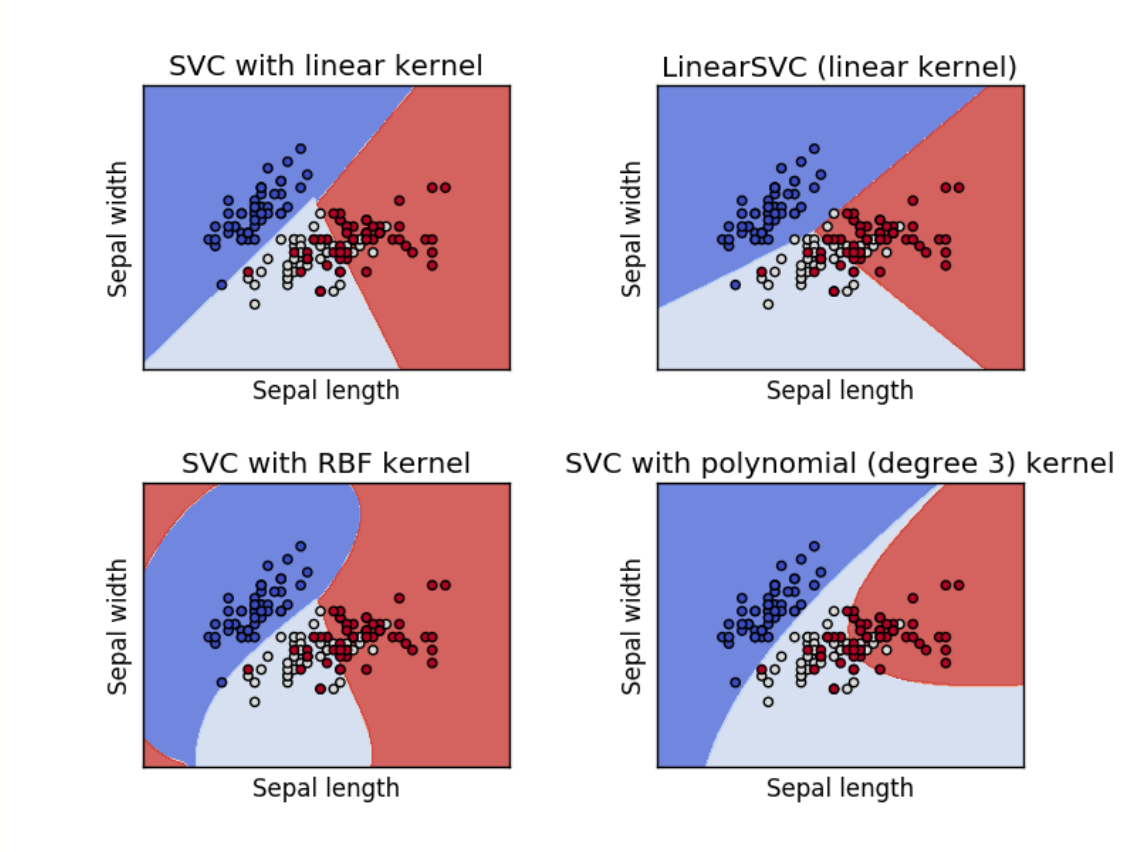
Support Vector Machines

- Non-Parametric model?

Read this article:- Demystifying SVMs.



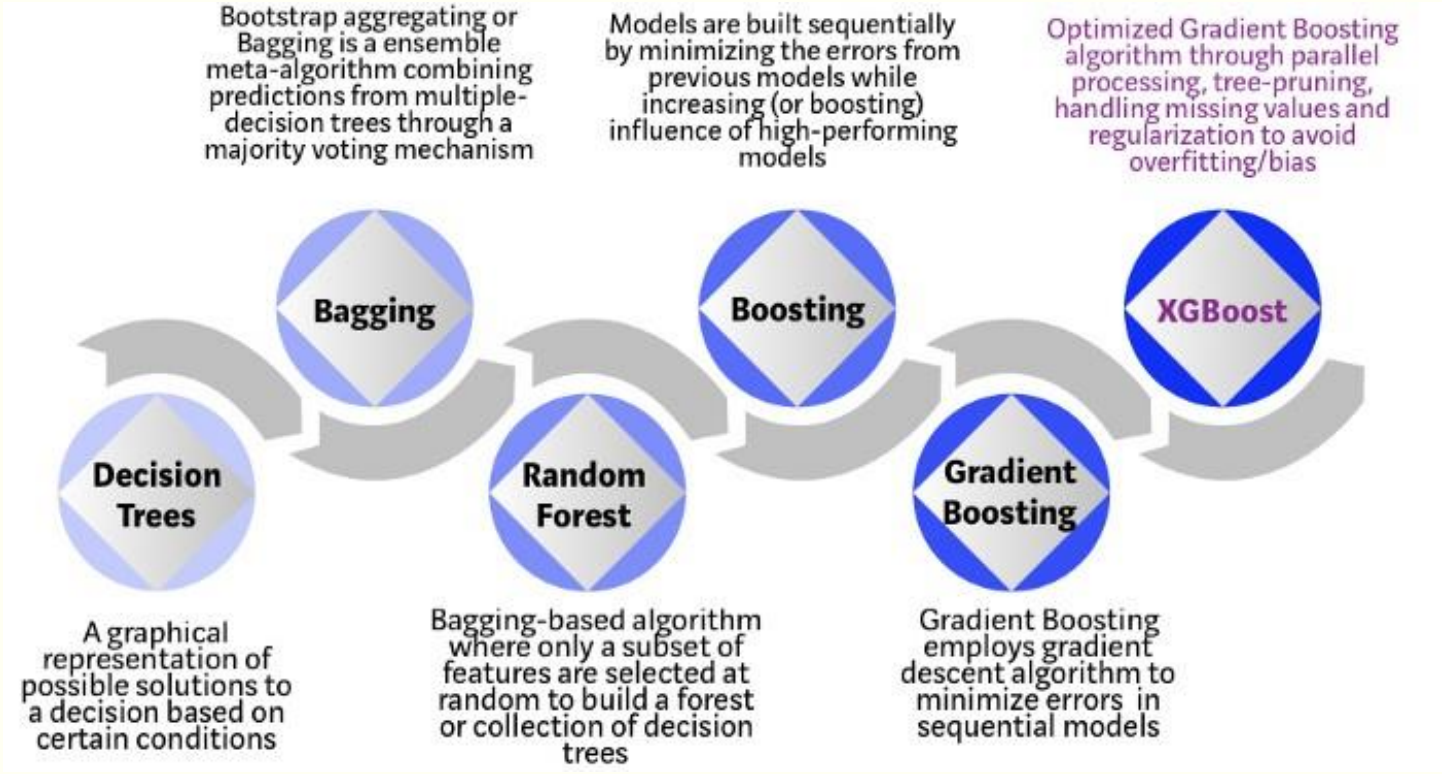
Source: Wikipedia



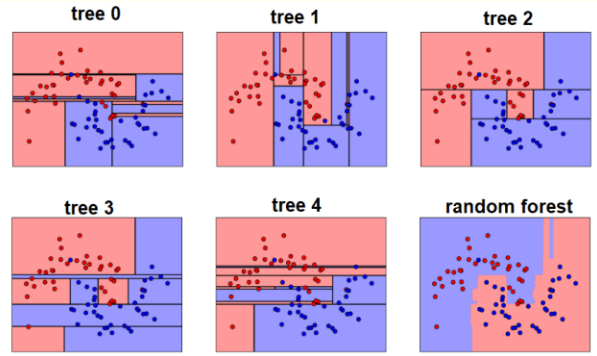
Source: Scikit Learn

Tree Based Classifiers

- Decision Trees:- Series of If-else statements
- Random Forest:- Ensembling several Decision Trees using bootstrap aggregation.
- XGBoost (***Long Live the Queen!***):- Boosting + many more cool things

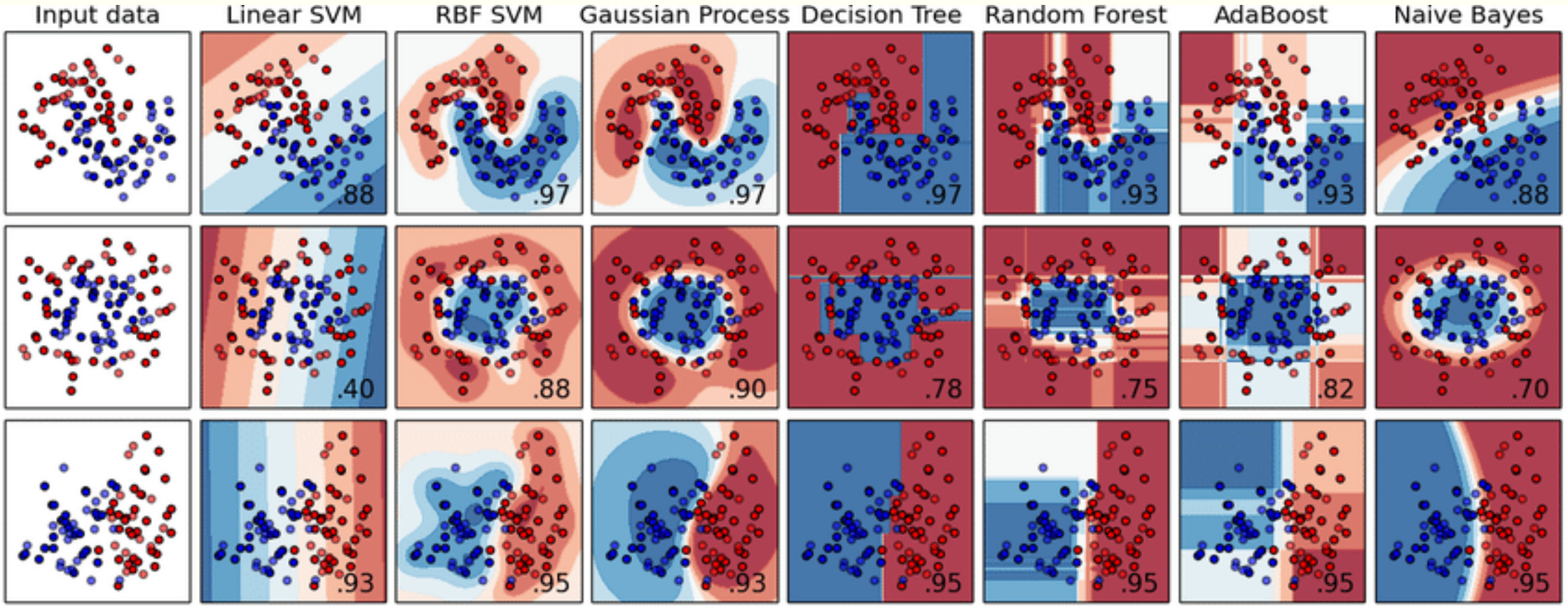


Source: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d



Source: Scikit Learn

Relative Comparison of Decision Boundaries



Source: Scikit Learn

Metrics used to judge a Classifiers Performance - I

- **Accuracy** = Correct Predictions / Total Predictions
- **Error** = Incorrect Predictions / Total Predictions
- **Sensitivity/Recall** refers to the true positive rate and summarizes how well the positive class was predicted.

  Sensitivity/Recall = TP / (TP + FN)

- **Specificity** is the complement to sensitivity, or the true negative rate, and summarizes how well the negative class was predicted.

  Specificity = TN / (FP + TN)

- Precision summarizes the fraction of examples assigned the positive class that belong to the positive class.

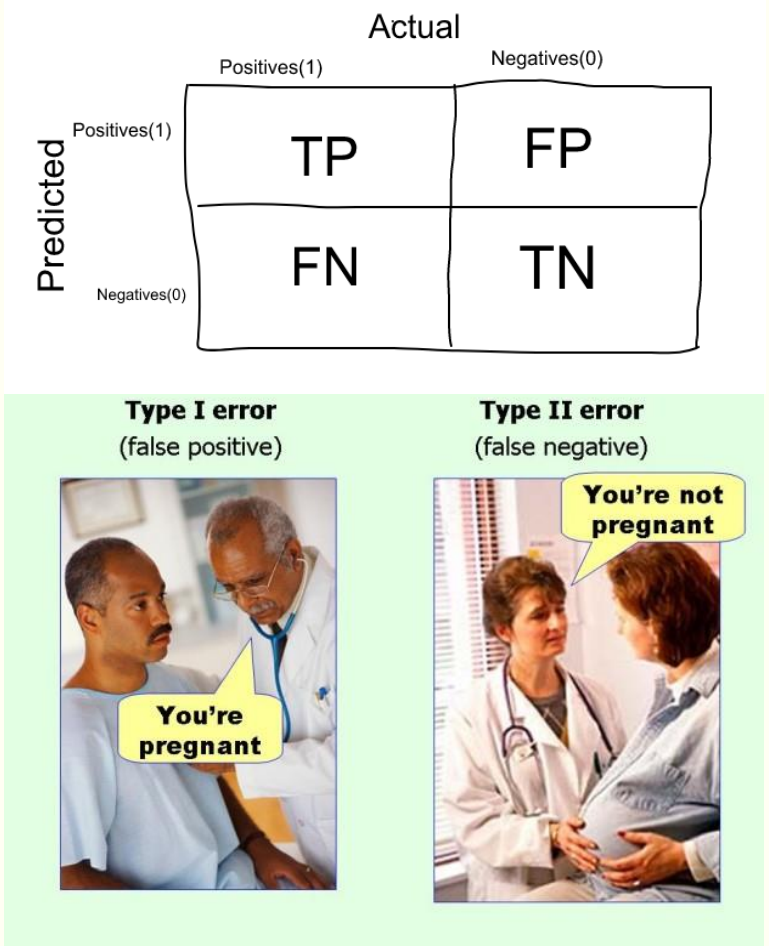
  Precision = TP / (TP + FP)

- Recall summarizes how well the positive class was predicted and is the same calculation as sensitivity.

  Recall = TP / (TP + FN)

- Precision and recall can be combined into a single score that seeks to balance both concerns, called the F-score and is a popular metric for imbalanced classification .

  F-Measure = (2 * Precision * Recall) / (Precision + Recall)

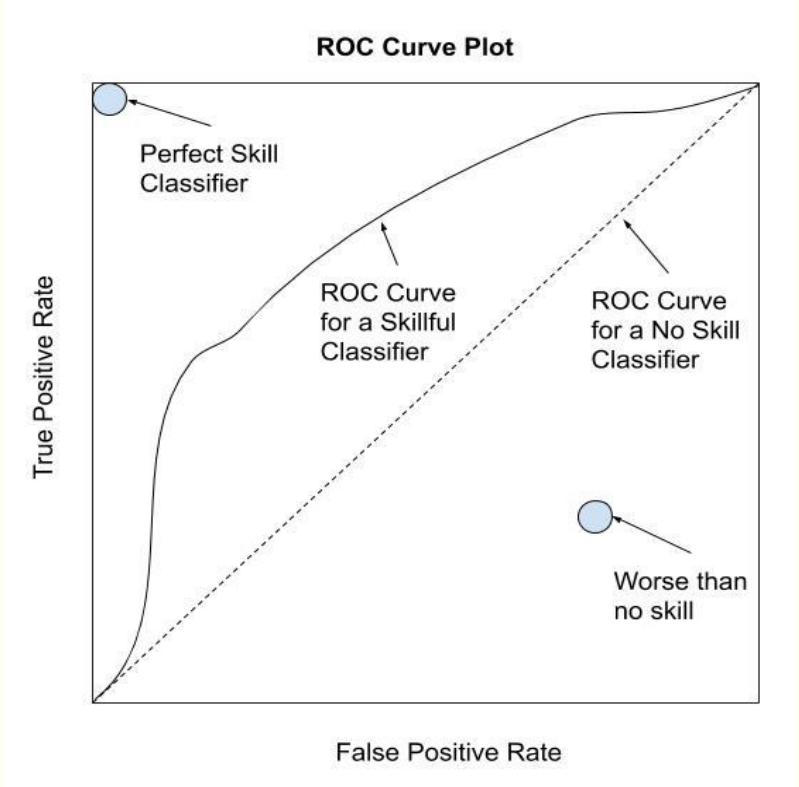


Source: TowardsDataScience

Metrics used to judge a Classifiers Performance - II

AUROC/ ROC (Receiver Operating Characteristic) Curve:-

- Compares Binary Classifiers based on their ability to discriminate between classes.
- A ROC curve is a diagnostic plot for summarizing the behavior of a model by calculating the false positive rate and true positive rate for a set of predictions by the model under different thresholds.
- Each threshold is a point on the plot and the points are connected to form a curve.
- The area under the ROC curve can be calculated and provides a single score to summarize the plot that can be used to compare models. A no skill classifier will have a score of 0.5, whereas a perfect classifier will have a score of 1.0.



Source: Machine Learning Mastery

Kohens Cappa Coefficient:-

- κ considers the possibility of the agreement occurring by chance. Many controversial debates surrounding this metric
- Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

References

- Machine Learning Mastery (Jason Brownlee)
- Machine Learning: A Probabilistic Perspective by Kevin Murphy
- 100-page Machine Learning Book by Andy Burkov
- Elements of Statistical Learning by Trevor, Hastie
- Artificial Intelligence: A Modern Approach by Russel & Norvig.