Garanti BBVA Technology

Classification Models

Week 3

Garanti BBVA Teknoloji

Ajanda

01 Classification Methods

02 Logistic Regression

03 Decision Trees

04 KNN

05 Performance Metrics

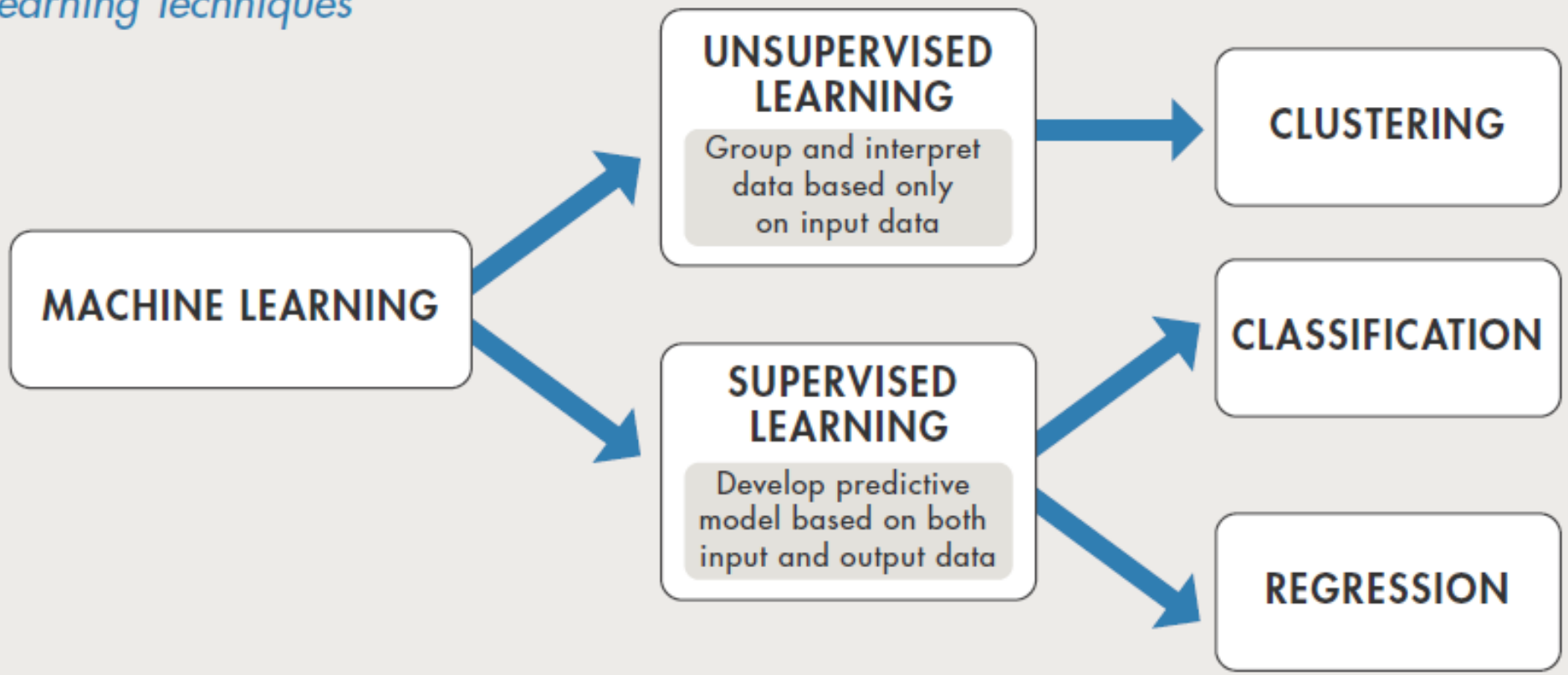
06 Practice on Mock Data

Garanti BBVA Teknoloji

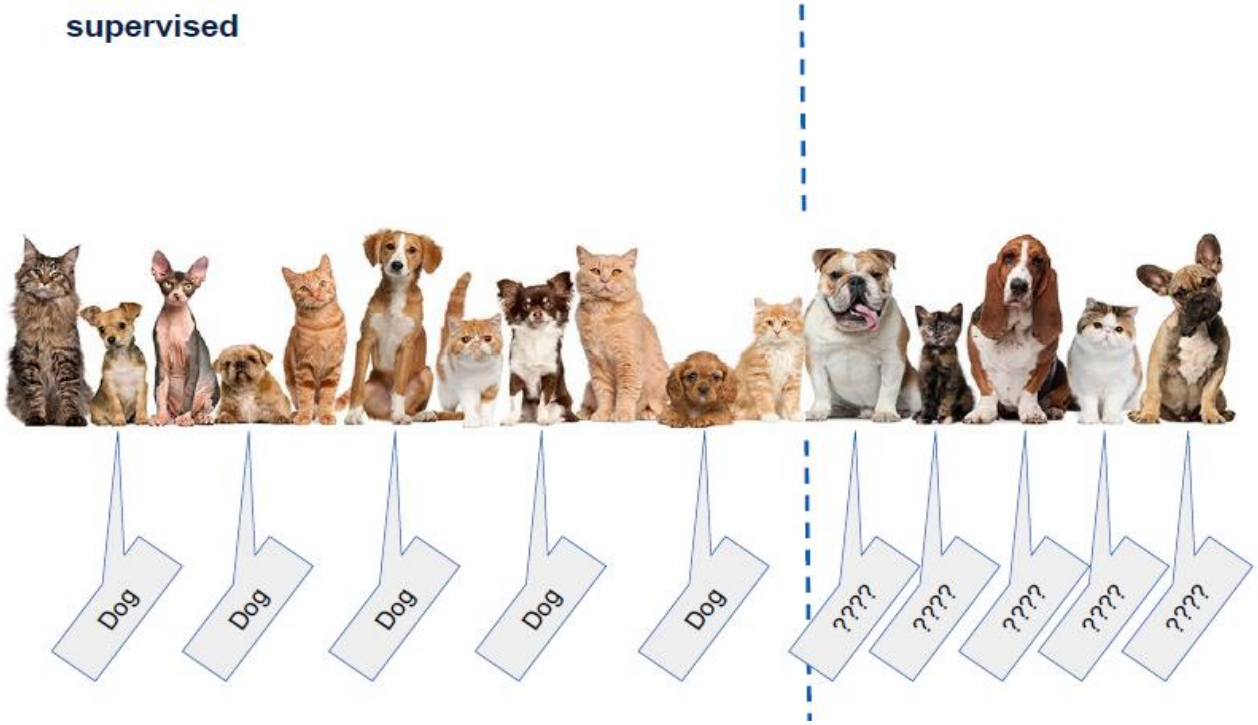
01

Classification Methods

Machine Learning Techniques



Supervised Learning



Training Dataset

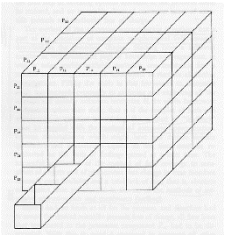
		Predictors (Features)				Labels (output Y)
		x_1	x_2	.....	x_p	Y
Samples	1					
	n					

n training samples or instances

p Predictors

Topology, Categories: Unveiling Compexity

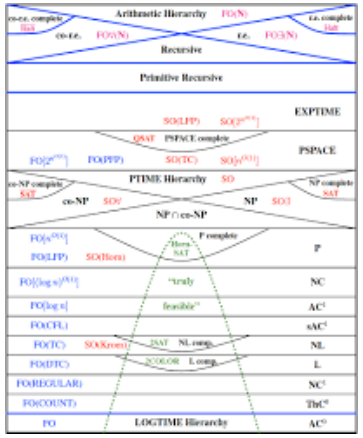
Category Theory



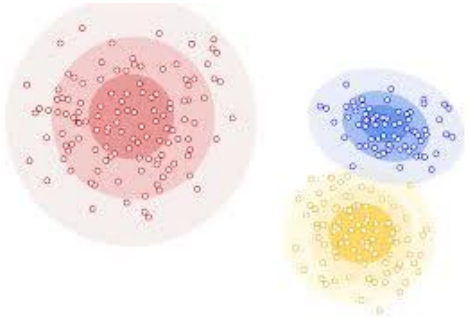
Morphological Analysis

Homologic Analysis

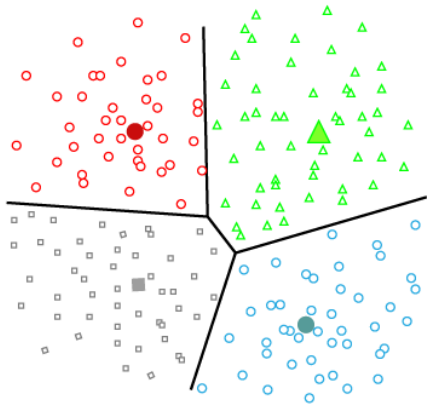
∞



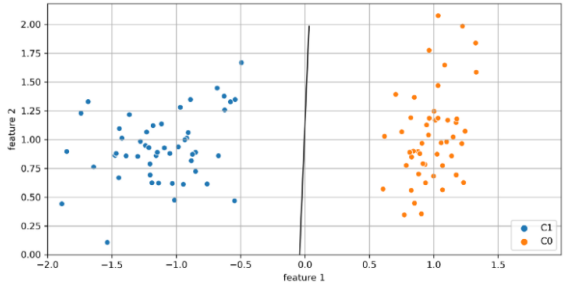
Computability Theory Topology



Clustering Topology



Multi Class Classification Topology



Binary Classification Topology

∞

Entropy D

Entropy C

Entropy B

Entropy A

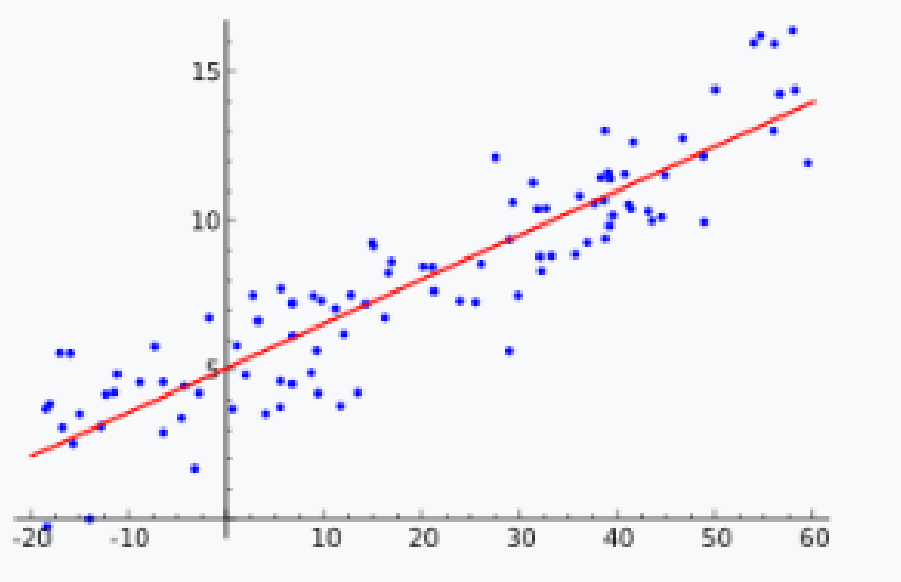
Problem

Garanti BBVA Teknoloji

02

Logistic Regression

What We Learned? -- Linear Regression



$$y=\beta_0+\beta_1 x$$

where

y: model estimation

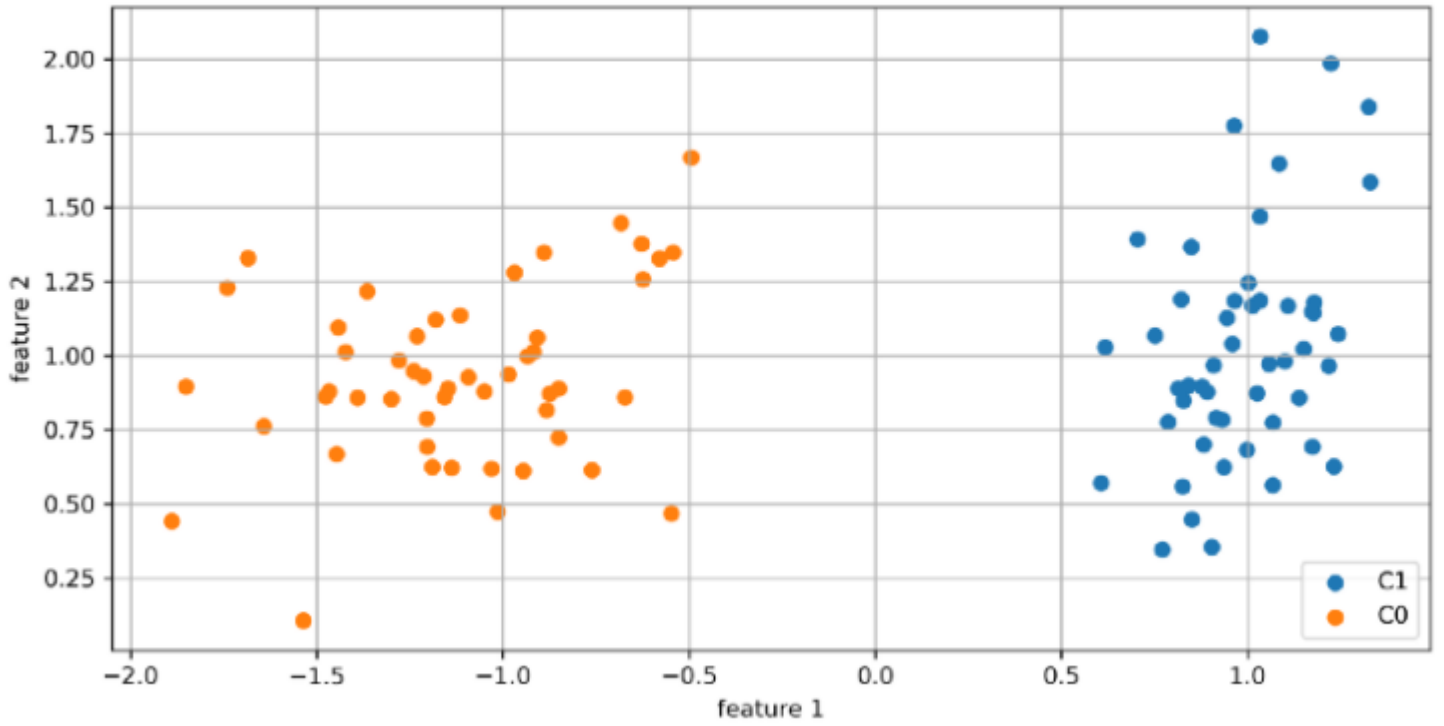
x : the input variable or predictor

B_0: intercept (x=0)

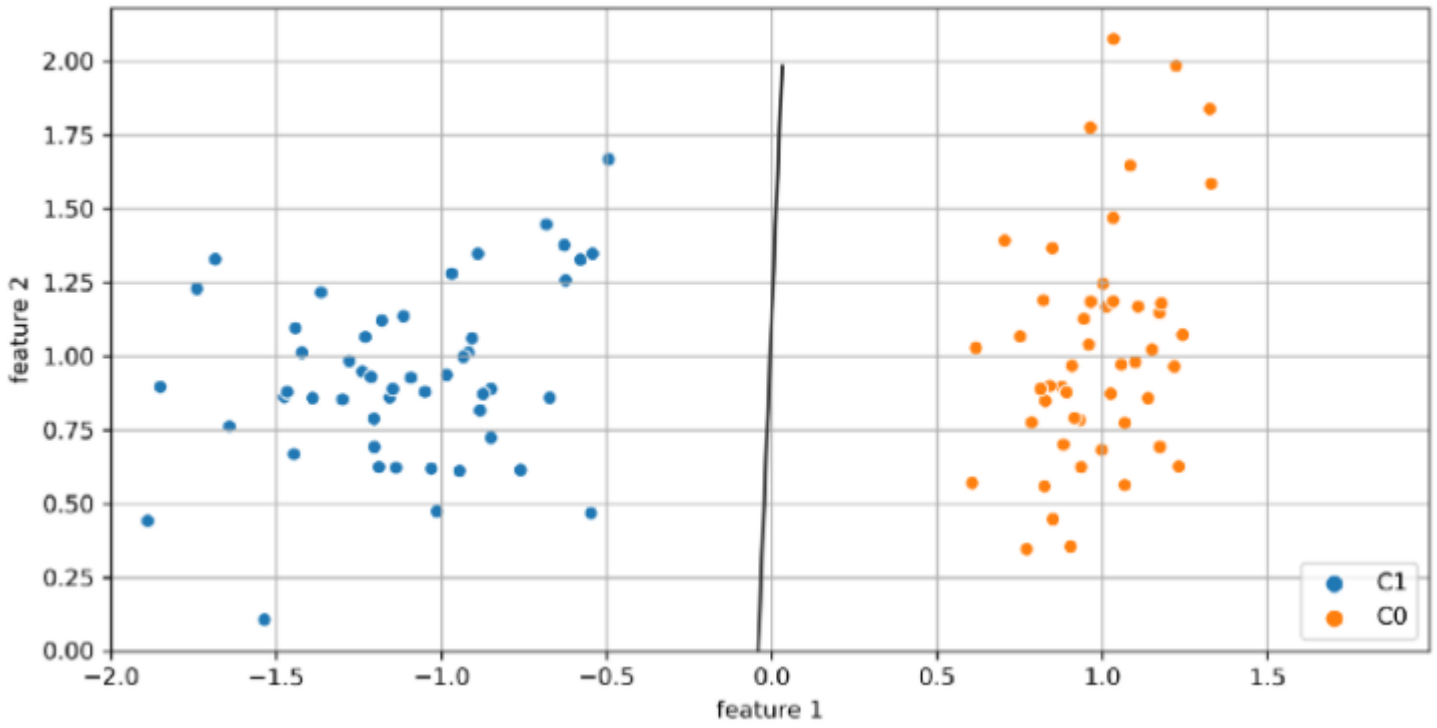
B_1: slope (coef of feature)

$$J(\beta_0,\beta_1)=\frac{1}{2m}\sum_{i=1}^{m}[y_i-(\beta_0+\beta_1 x_i)]^2$$

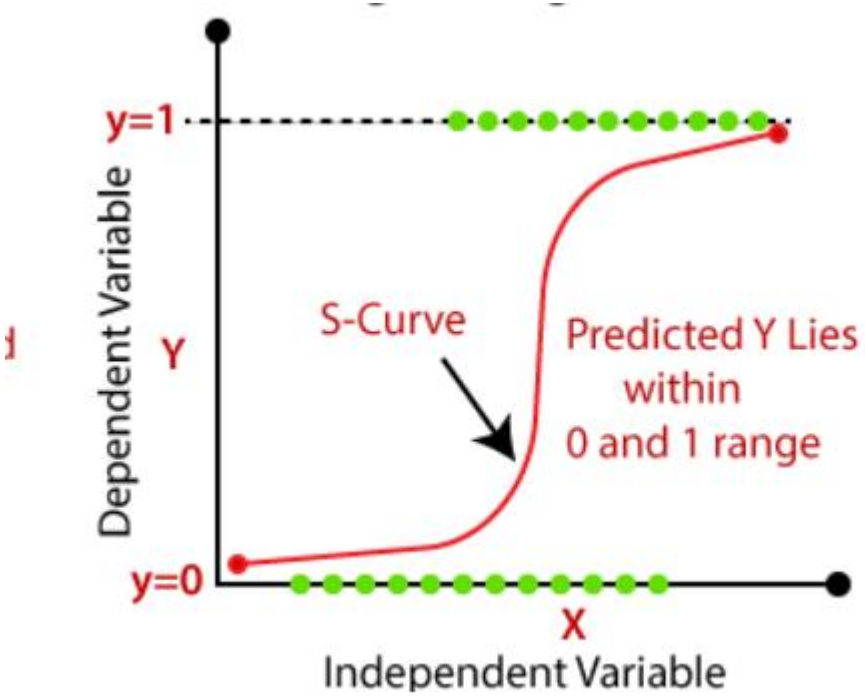
Logistic Regression with 2 Feature



Decision Boundary



Sigmoid Function



- Binary & Categorical target
- A form of linear regression
- Sigmoid Function

$$y = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Assumptions & Advantages

Assumptions

- Linear relationship between features and target
- Minimum or no relationship between features
- Relationship between one feature and target is indepented of other feature and target
- No extreme Outliers
- Sample size is large enough

Advantages

- Results interpretability
- Ease of use
- Low computational cost

Overfitting & Regularization

- Lasso (L1) Regularization

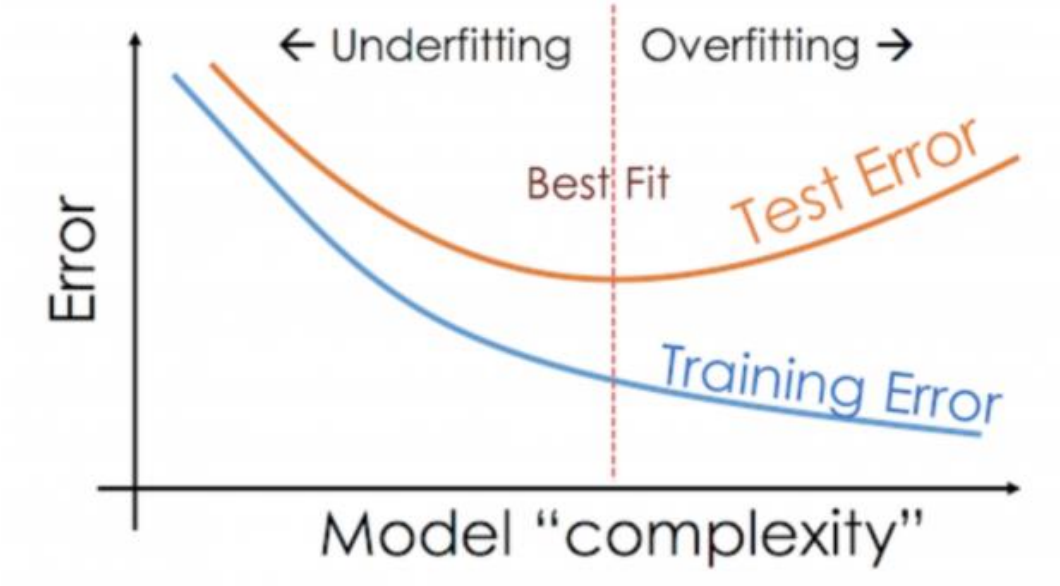
$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Cost function

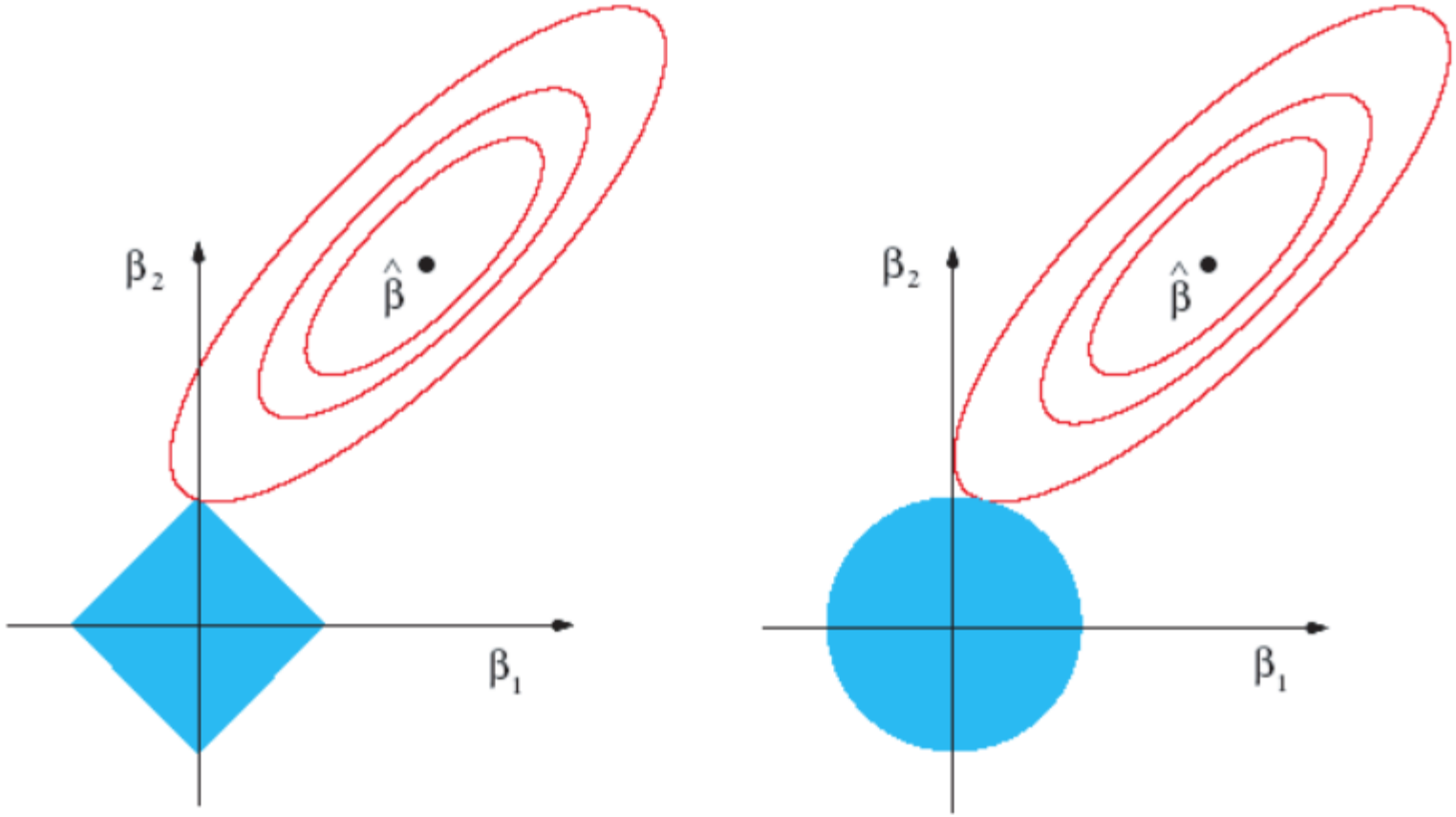
- L2: Ridge (L2) Regularization

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Cost function

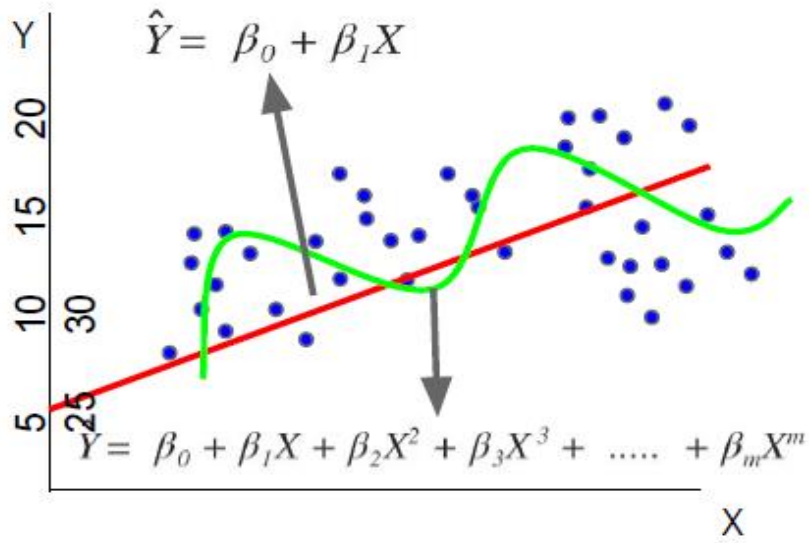


Lasso vs Ridge Regularization

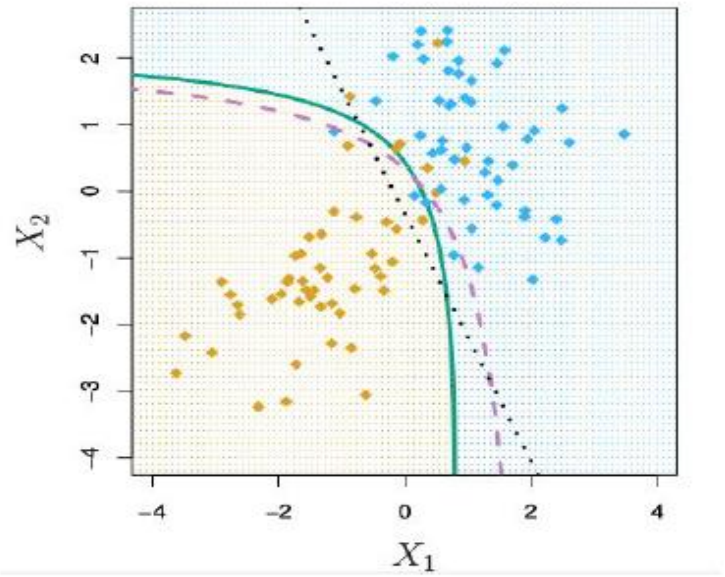


Summary

Regression



Classification



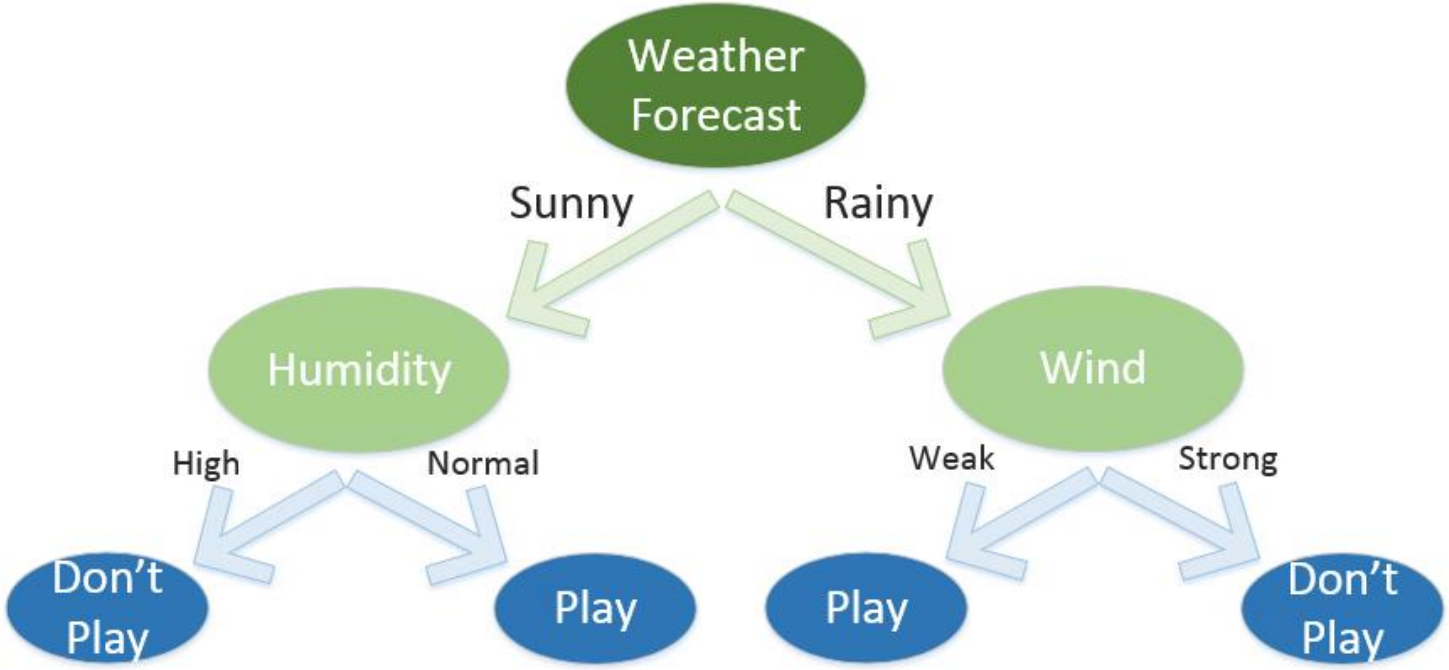
Garanti BBVA Teknoloji

03

Decision Trees

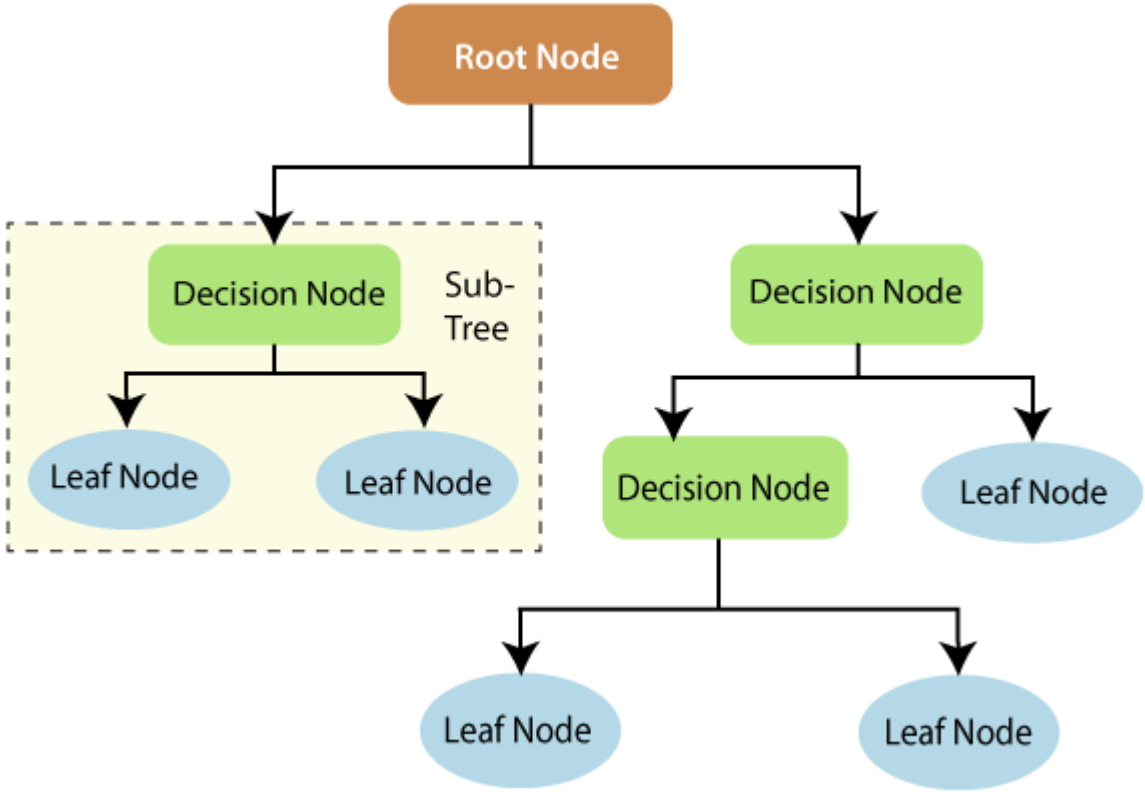
Decision Trees

- One of the most widely used and practical methods for inductive inference
- **CARTs** can be used for classification or regression problems (mostly classification)
- Weaknesses are relatively minor and can be largely avoided.



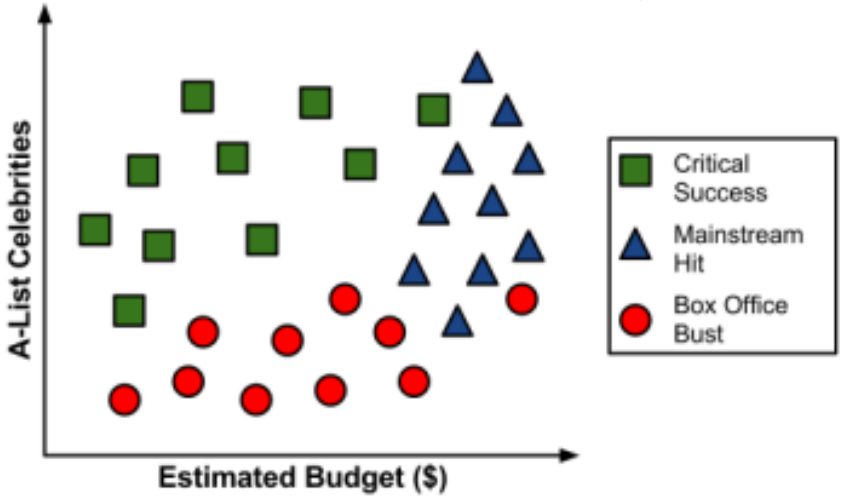
Structure of a Decision Tree

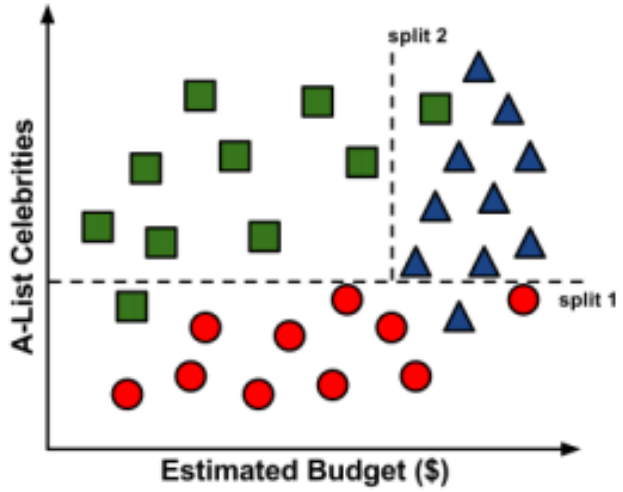
- A flow-chart-like tree structure
- **Decision node** denotes a test on an attribute
- **Branch** represents an outcome of the test
- **Leaf nodes** represent class labels or class distribution



Divide and Conquer

- Decision trees are built using a heuristic called recursive partitioning
- It splits the data into smaller and smaller subsets of similar classes
- Divide and conquer the nodes until a stopping criterion is reached
  - All (or nearly all) of the examples at the node have the same class
  - There are no remaining features to distinguish among examples
  - The tree has grown to a predefined size limit



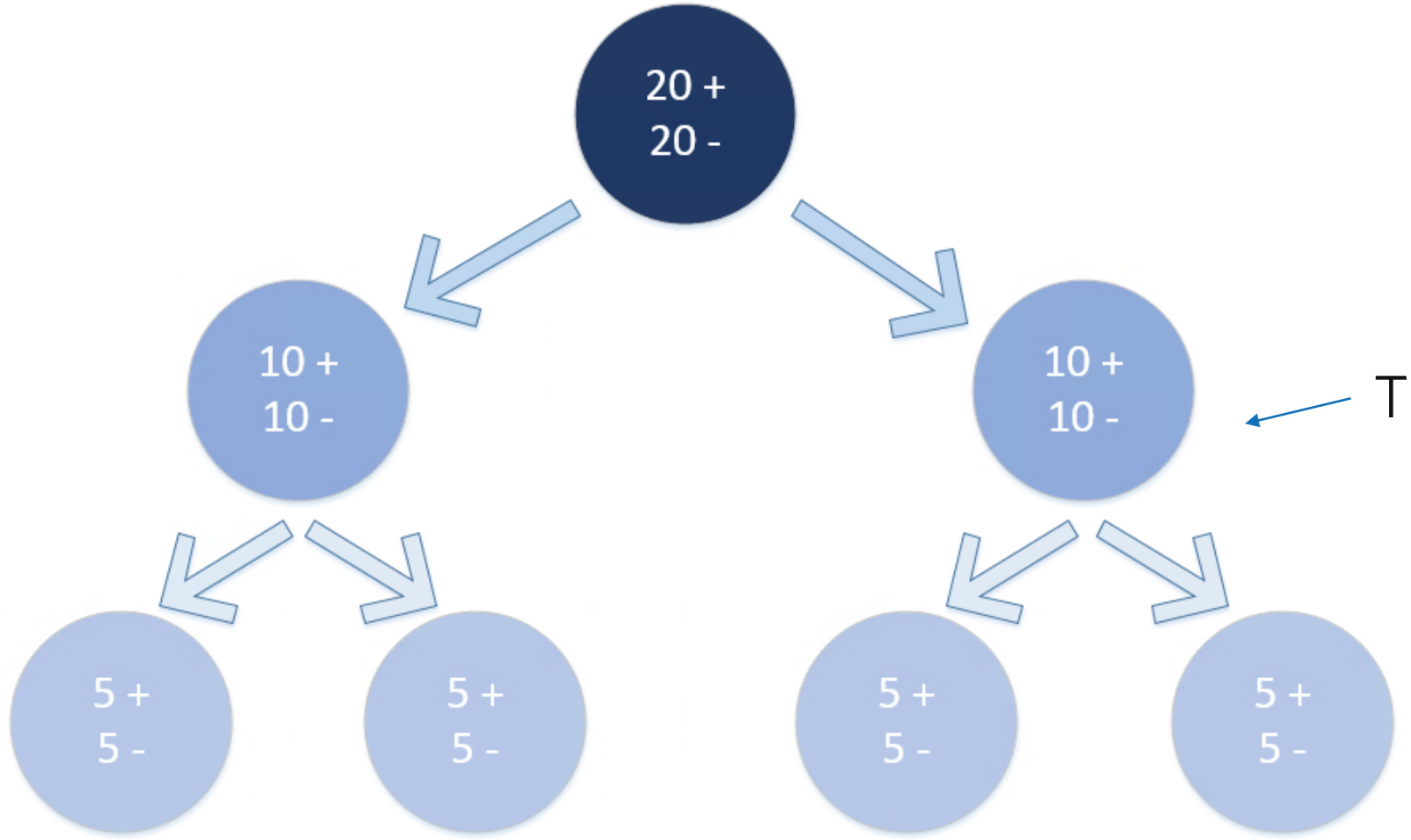


Divide and Conquer – Classification Using Decision Trees and Rules, Author : Brett Lantz

Learning the Tree

- Many algorithms: ID3, J48, C4.5, C5.0
- Algorithmic issues:
  - **Determine how to split the records**
    - How to specify the attribute test condition?
    - How to determine the best split?
  - **Determine when to stop splitting**
  - **How to classify a leaf node**
    - Assign the majority class
    - If leaf is empty, assign the class with the most popular label

Which Feature is Best for Splitting?



This is bad splitting!

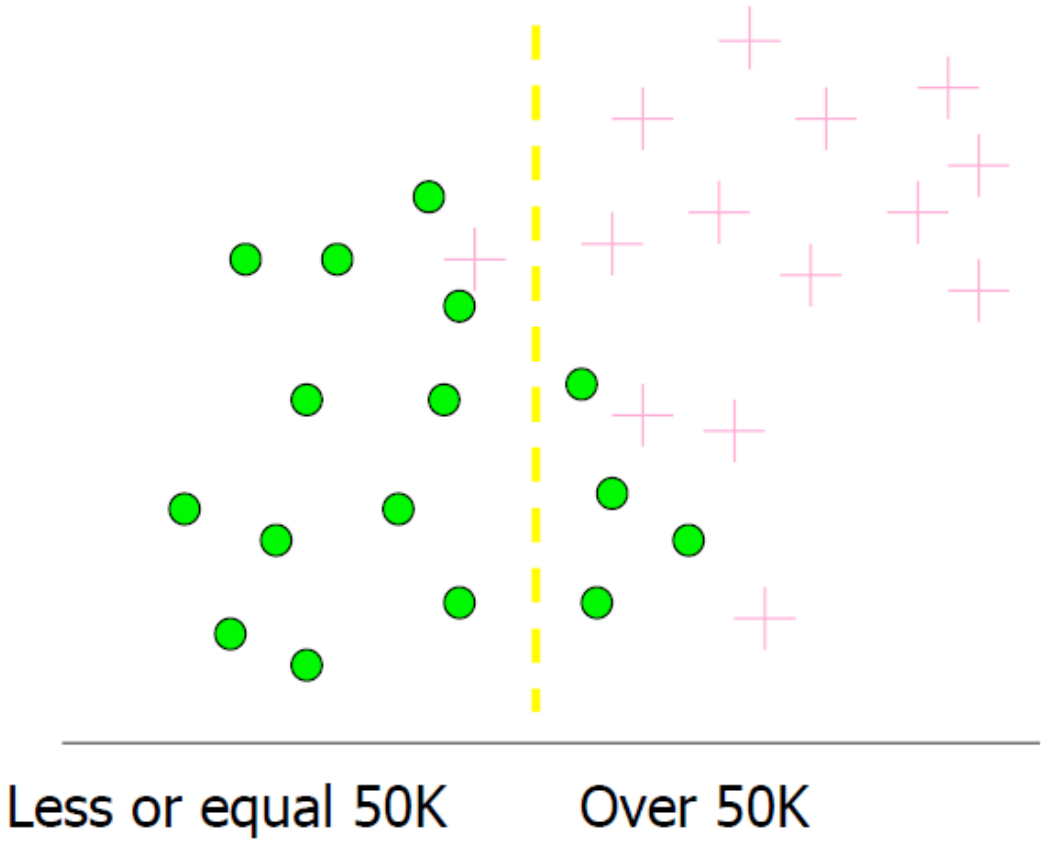
Ideal:

18 / 2

+9 / -1 +9 / -1

Determining the Best Split

Split over whether Balance exceeds 50K

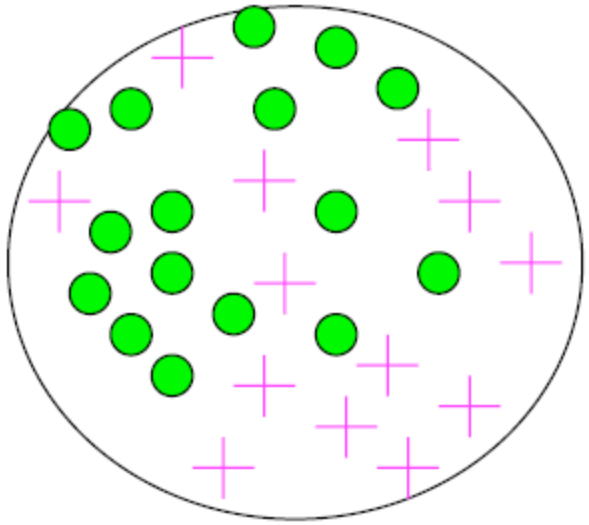


Split over whether applicant is employed

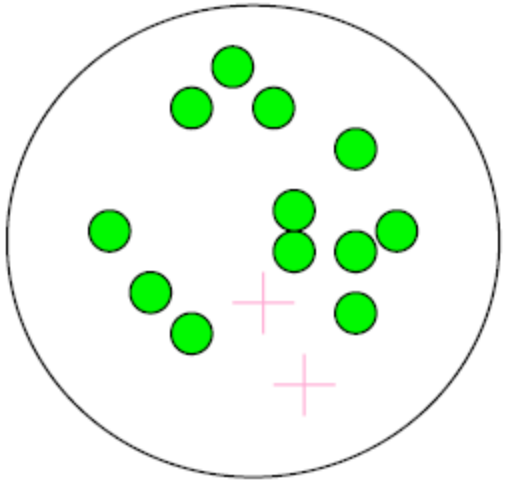
Impurity

We need a measure of impurity for class label distribution on a node

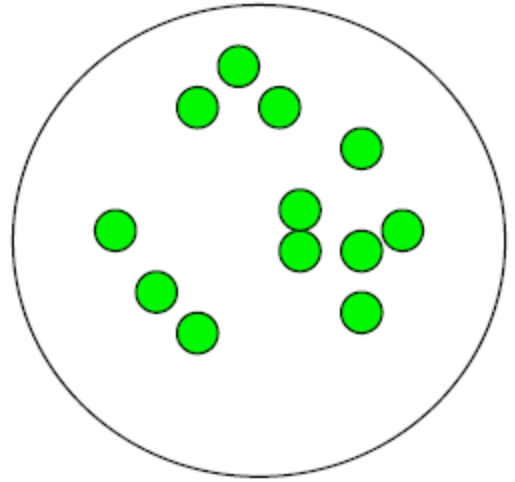
Very impure group



Less impure

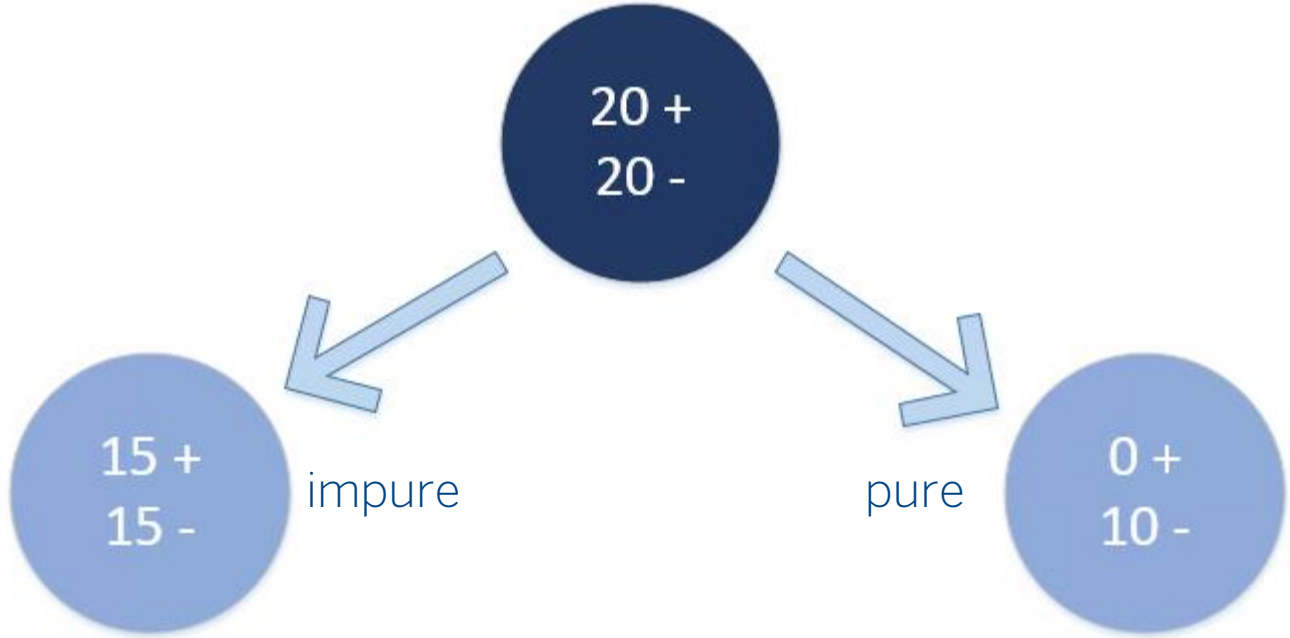


Minimum impurity



Common Measures of Impurity

- Entropy
- Gini Index
- Misclassification Error



If the distribution is **less uniform, the node is purer.**

Information & Entropy

Information : Amount of uncertainty (amount of surprise) in the outcome

- Flipping coin
- Rolling a dice

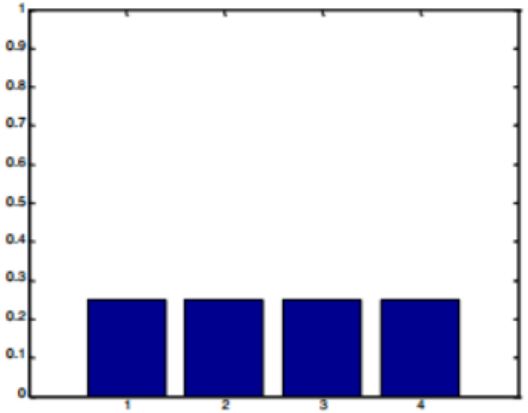
$$I(X) = \log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

- If the probability of an event is small and it happens, the information is large
  - Flipping «head» on a coin > $I = -\log_2 1/2 = 1$
  - Rolling «6» on a dice > $I = -\log_2 1/6 = 2.58$

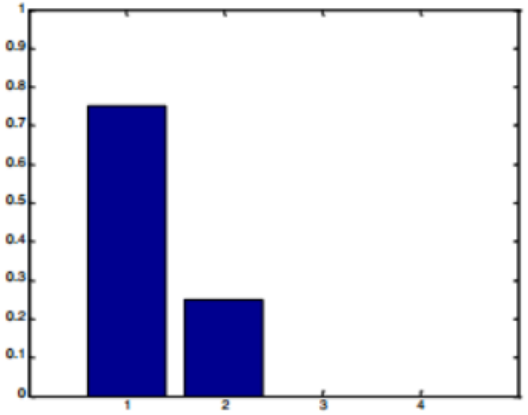
Information & Entropy

Entropy : The expected amount of information when observing the output of a random variable X:

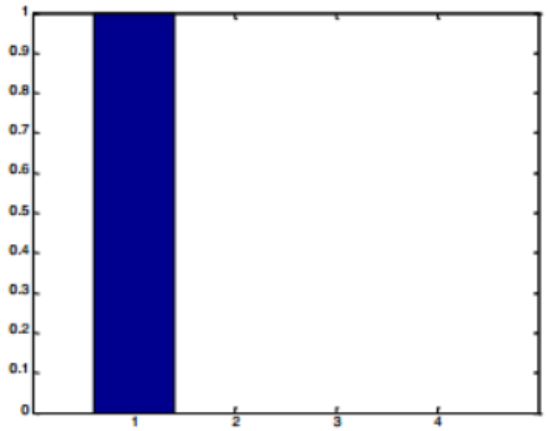
$$H(X) = E(I(X)) = \sum_i p(x_i) I(x_i) = -\sum_i p(x_i) \log_2 p(x_i)$$



H(x) = .25 log 4 + .25 log 4 + .25 log 4 + .25 log 4 = log 4 = 2 bits



H(x) = .75 log 4/3 + .25 log 4 ≈ .8133 bits



H(x) = 1 log 1 = 0 bits

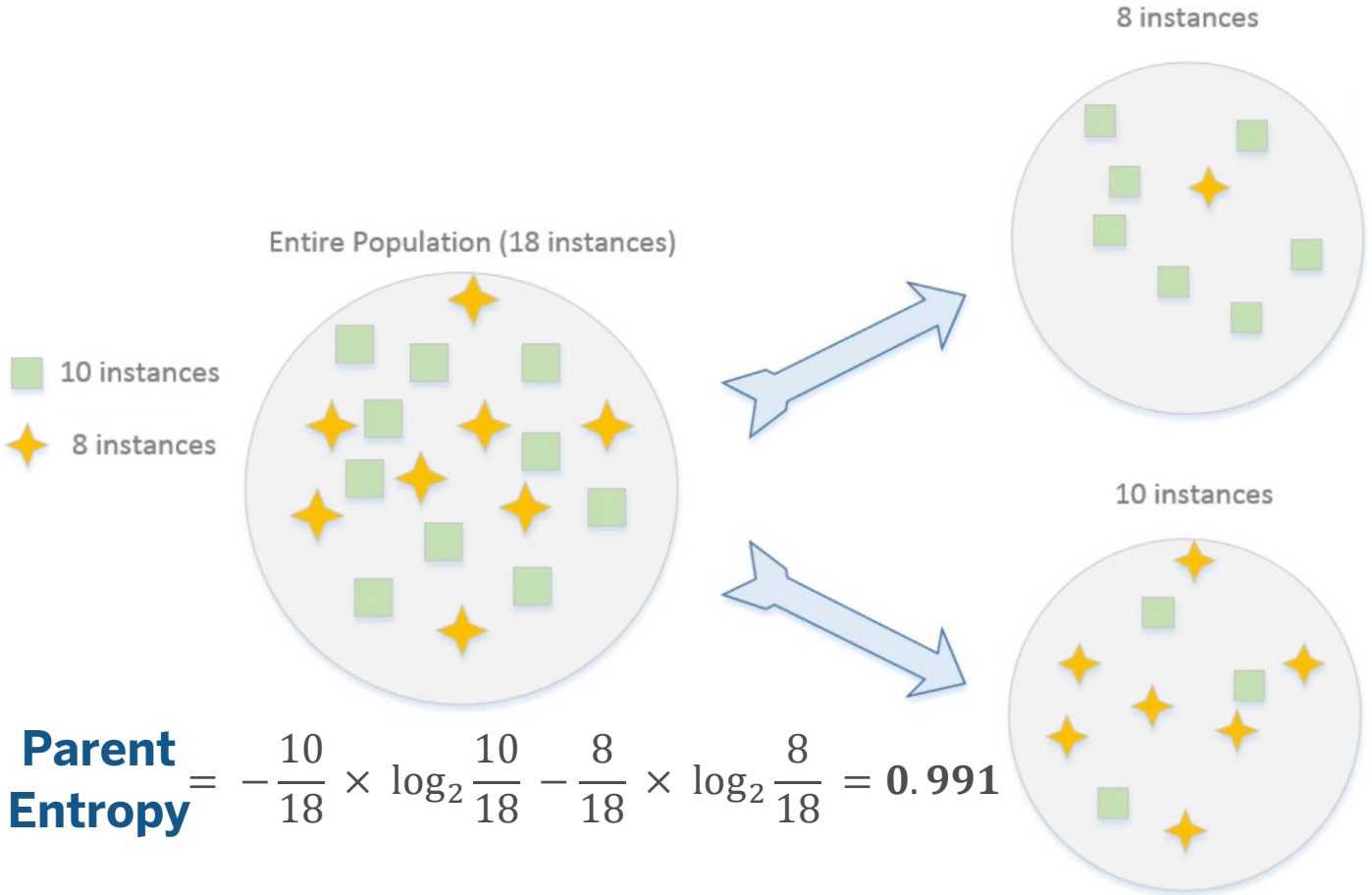
Information Gain

Information Gain : Reduction in uncertainty by knowing Y

$$IG(X,Y)=H(Y) - H(Y|X)$$

- Information gain tells us the importance of the given feature
- Determines the most useful one for discriminating between the classes to be learned.

Calculation of Information Gain



Child Entropy

$$= -\frac{7}{8} \times \log_2 \frac{7}{8} \; -\frac{1}{8} \times \log_2 \frac{1}{8} = \mathbf{0.544}$$

Child Entropy

$$-\frac{3}{10} \times \log_2 \frac{3}{10} \; -\frac{7}{10} \times \log_2 \frac{7}{10} = \mathbf{0.881}$$

Parent Entropy

$$= -\frac{10}{18} \times \log_2 \frac{10}{18} - \frac{8}{18} \times \log_2 \frac{8}{18} = \mathbf{0.991}$$

Weighted Average Entropy of Children

$$= \left(\frac{10}{18} \times \mathbf{0.881}\right) + \left(\frac{8}{18} \times \mathbf{0.544}\right) = \mathbf{0.731}$$

Information Gain :

$$\mathbf{0.991 - 0.731 = 0.26}$$

Other Impurity Measures

Gini :

$$GINI(t) = 1 - \sum_i p(i \mid t)^2$$

$p(i \mid t)$ is the relative frequency of class *i* at node *t*

Misclassification Error :

$$\text{Classification error}(t) = 1 - \max_i p(i \mid t)$$

+	**2**
-	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = $1 - (2/6)^2 - (4/6)^2 = 0.444$

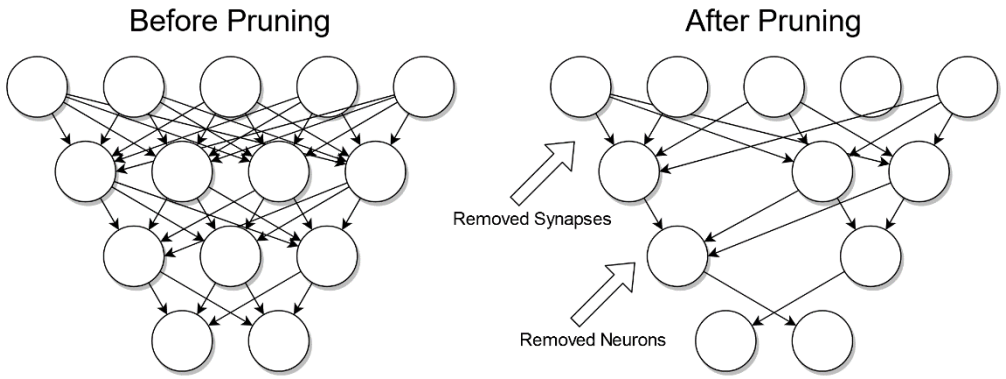
Entropy = $-(2/6)\log_2(2/6) - (4/6)\log_2(4/6) = 0.92$

Error = $1 - \max(2/6, 4/6) = 1 - 4/6 = 1/3$

Generalization of Decision Trees

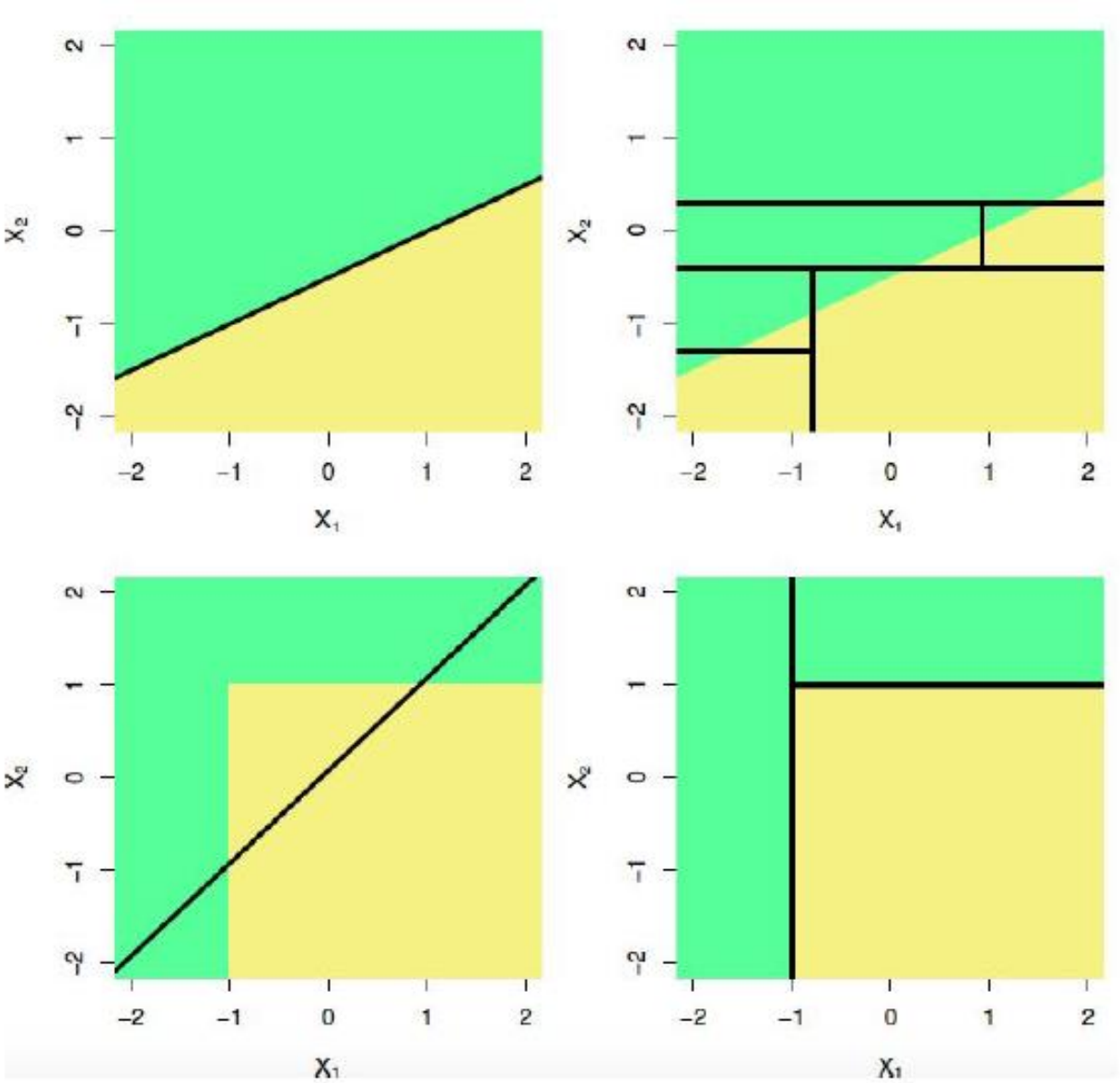
Pruning: Replacing a subtree with a leaf with the most common classification in the subtree.

- Pre Pruning (Early Stopping Rule) : Stop growing the tree before it perfectly classifies the training data.
- Post Pruning : Grow full tree, then prune.



Decision Boundaries (Lineer vs Trees)

- Decision boundary is parallel to axes because test condition involves a single feature at a time
- Tree models better reflect nonlinear relationships between feature and target.
- if the relationship is expected to be linear, linear models might be better.

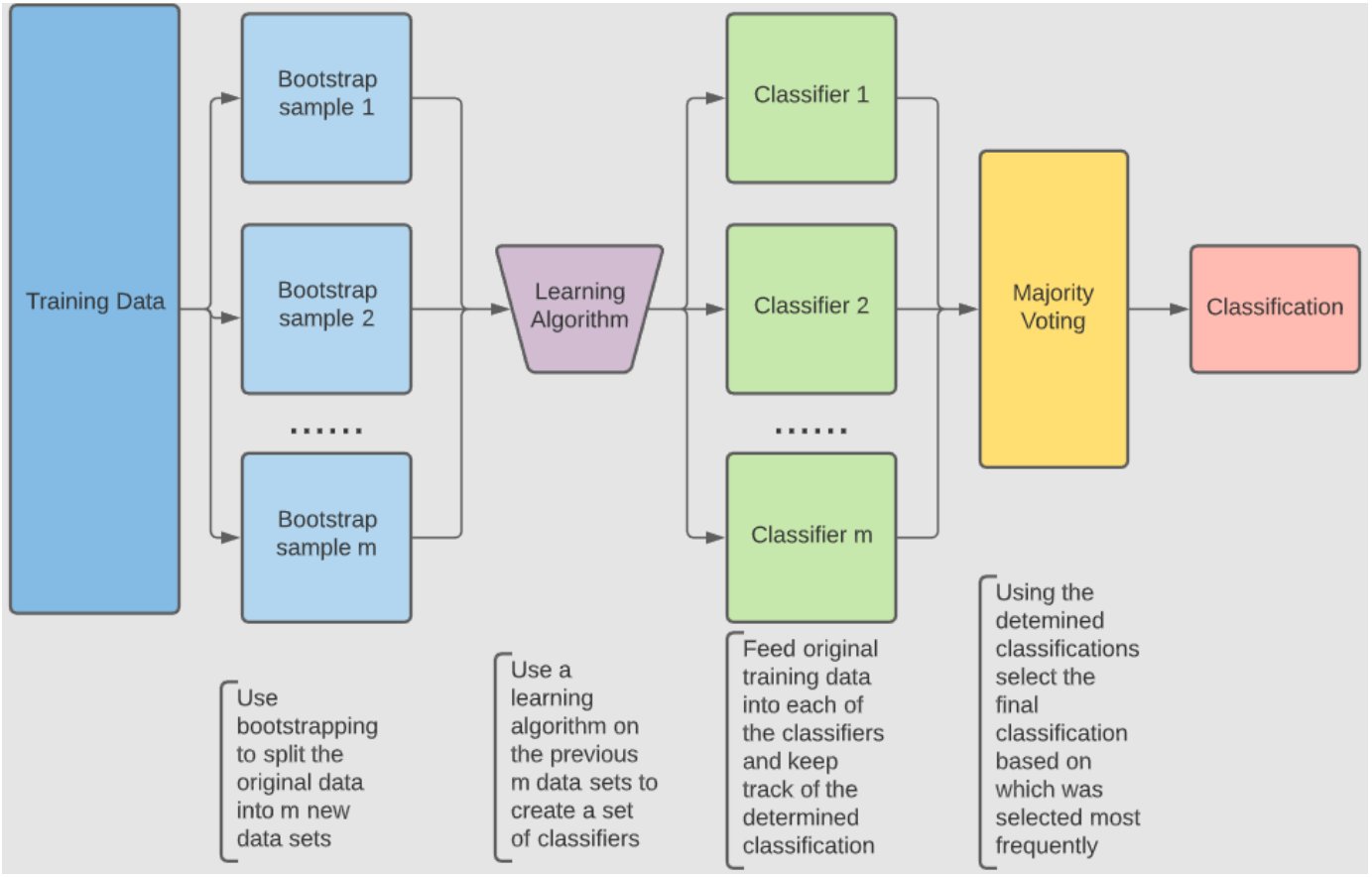


An Introduction to Statistical Learning. Figure 8.7, Authors :Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Bagging (Bootstrap Aggregation)

- Random Forest

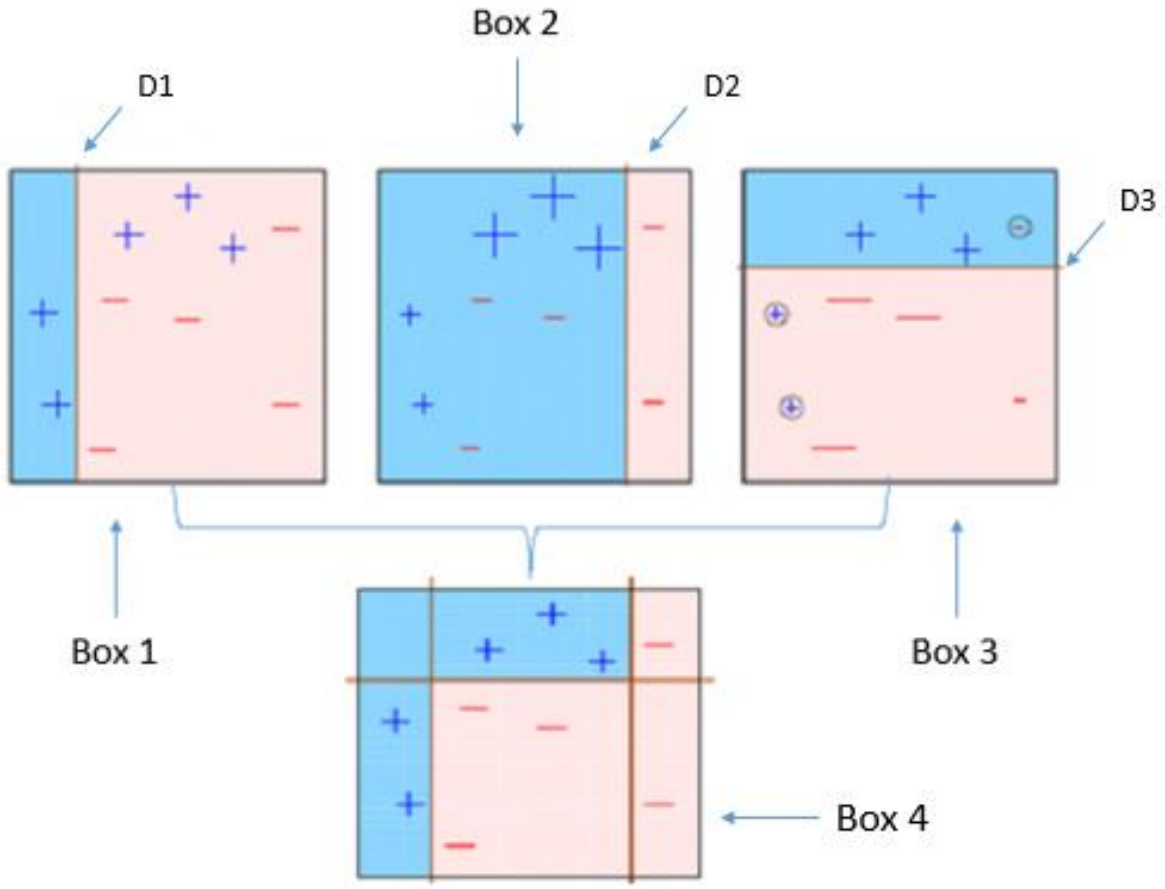
General procedure that can be used to **reduce the variance** for the decision tree algorithm that have high variance.



Boosting

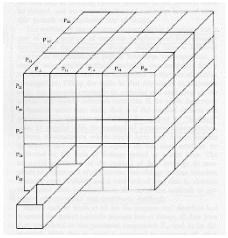
- Adaptive Boosting (AdaBoost)

General ensemble method that creates a strong classifier from a number of **weak classifiers**.



Topology, Categories: Unveiling Compexity

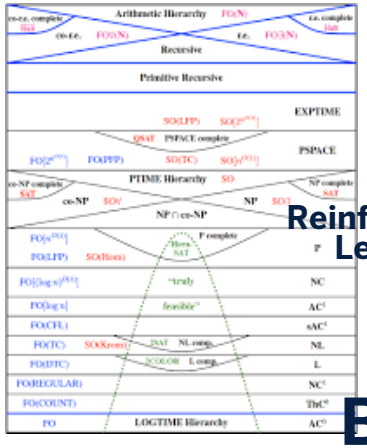
Category Theory



Morphological Analysis

Homologic Analysis

∞



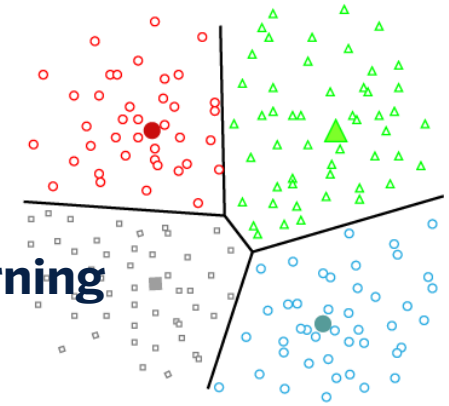
Computability Theory Topology

Boosting

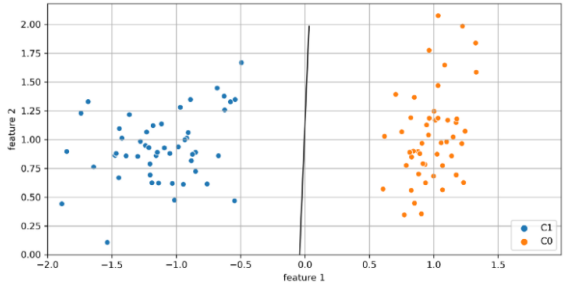
Reinforcement Learning

Deep Learning

Bagging



Multi Class Classification Topology



Binary Classification Topology

Clustering Topology

Entropy D

Entropy C

Entropy B

Entropy A

Problem

∞

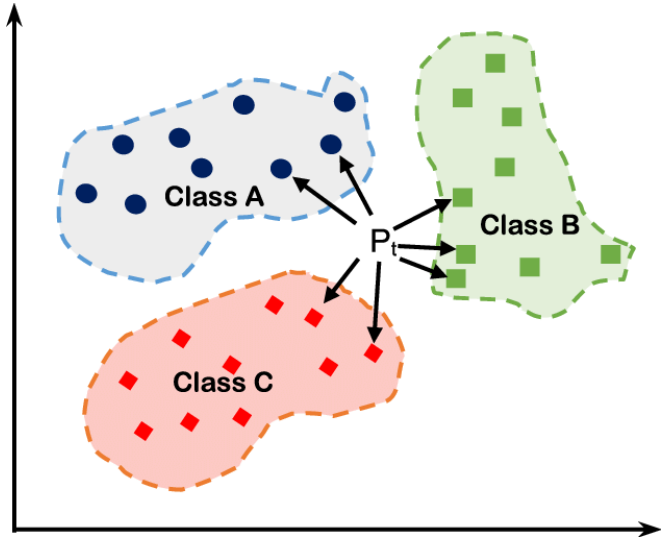
Garanti BBVA Technology

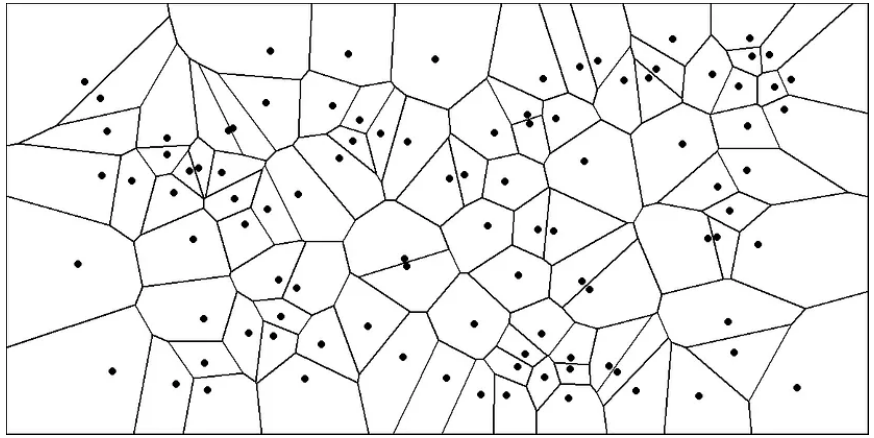
04

K Nearest Neighbors

k Nearest Neighbours

- One of the easy to implement non-probabilistic supervised learning algorithms.
- **KNN** can be used for classification or regression problems (mostly classification)
- Algorithm assumes that similar things exist in close proximity. It categorizes objects based on the classes of their nearest neighbours in the dataset.

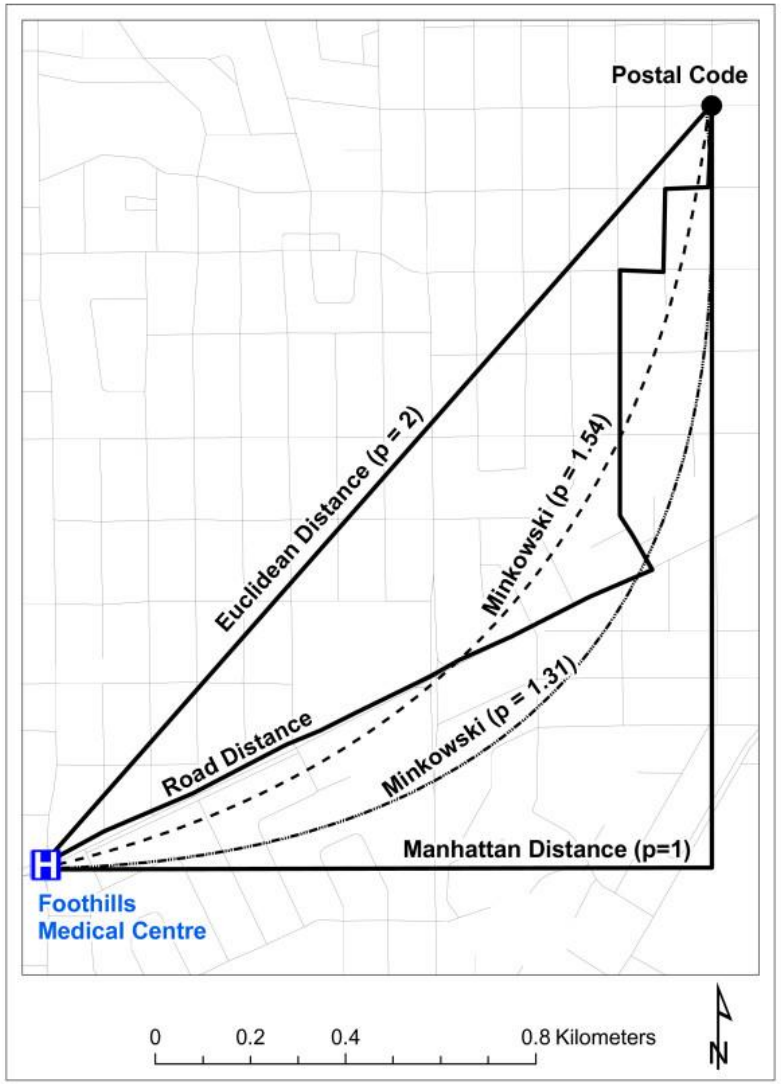




1-NN Pattern : Voronoi Diagrams

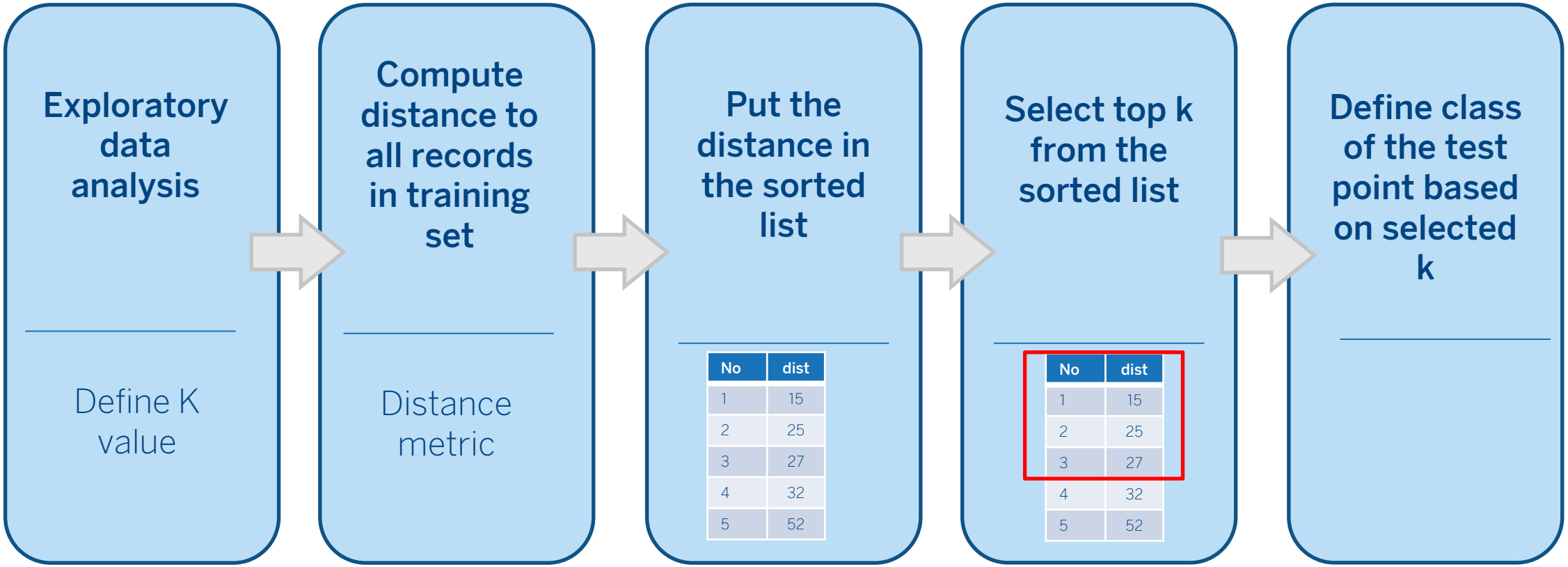
Measure of Distance in Data Mining

- Euclidean $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
- Manhattan $|x_1 - x_2| + |y_1 - y_2|$
- Minkowski $d(p, q) = (\sum_{i=1}^{n} |p_i - q_i|^c)^{\frac{1}{c}}$



Future Reading for Different Distance Metrics

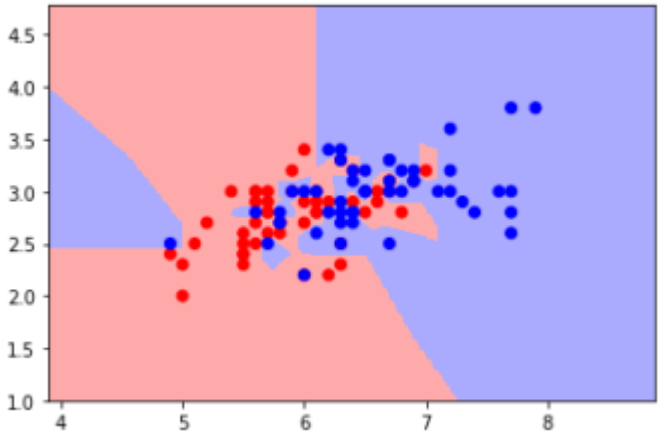
k Nearest Neighbours



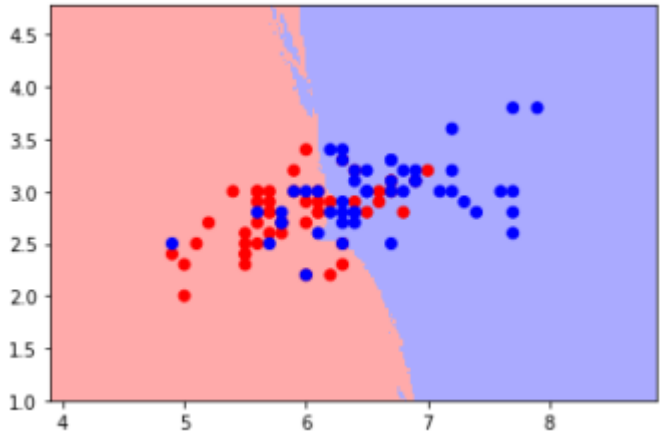
Visualization of KNN Algorithm

Choosing the k Value : Variance & Bias

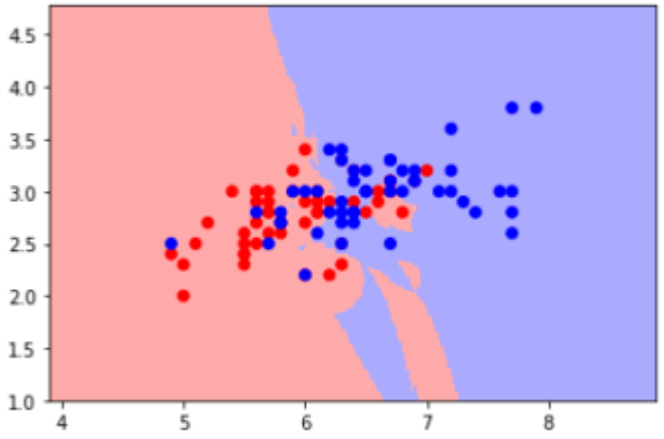
1



2

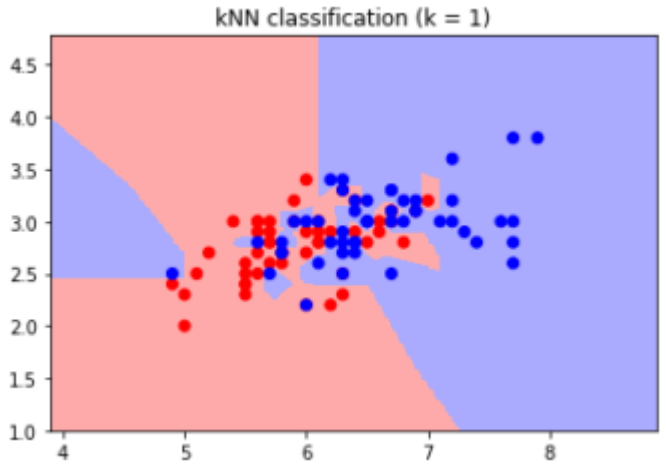


3

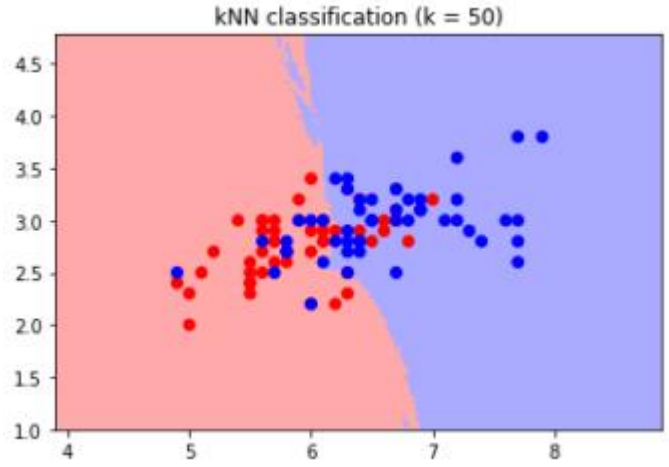


Choosing the k Value : Variance & Bias

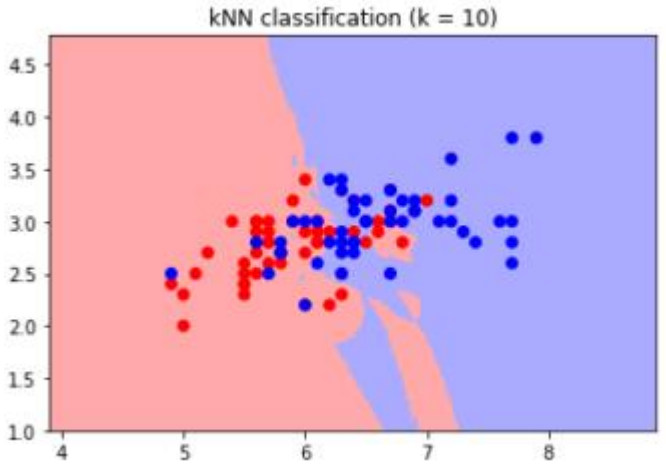
1



2



3

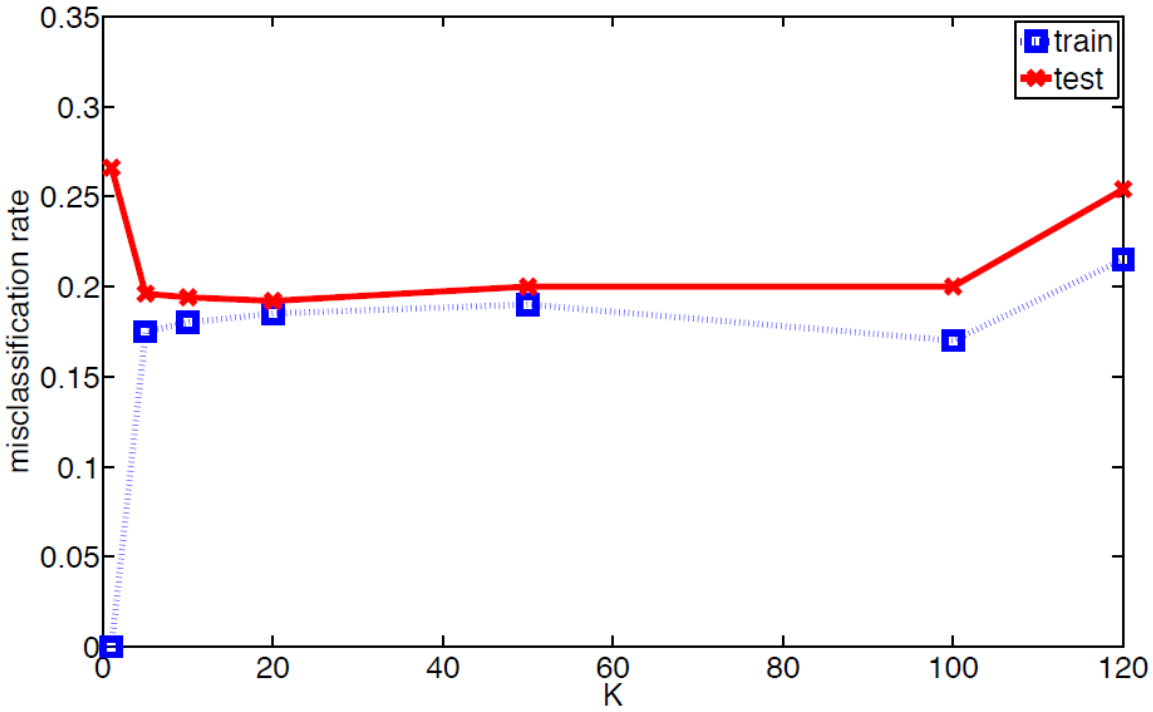


2 > 3 > 1

- **Small *k* values** cause **high variance** and an unstable output
- **Large *k* values** lead to **high bias** by classifying everything as the more probable class

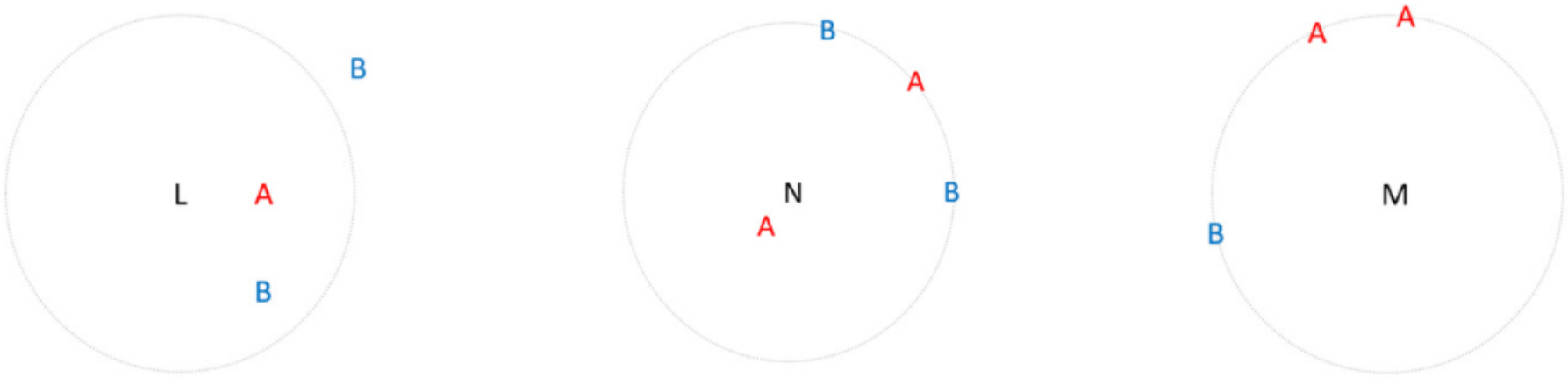
Choosing the k Value : Error Rate

- Rule of thumb is using **sqrt(N)** as k
- Or, iterate with different k and interpret the error rate



$$\mathrm{err}(f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$$

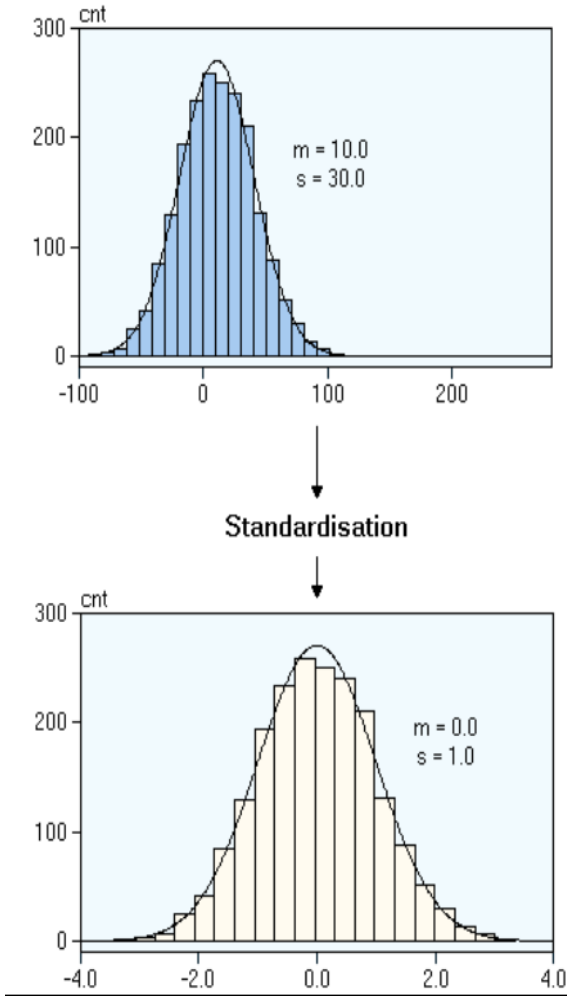
Choosing the k Value : Ties

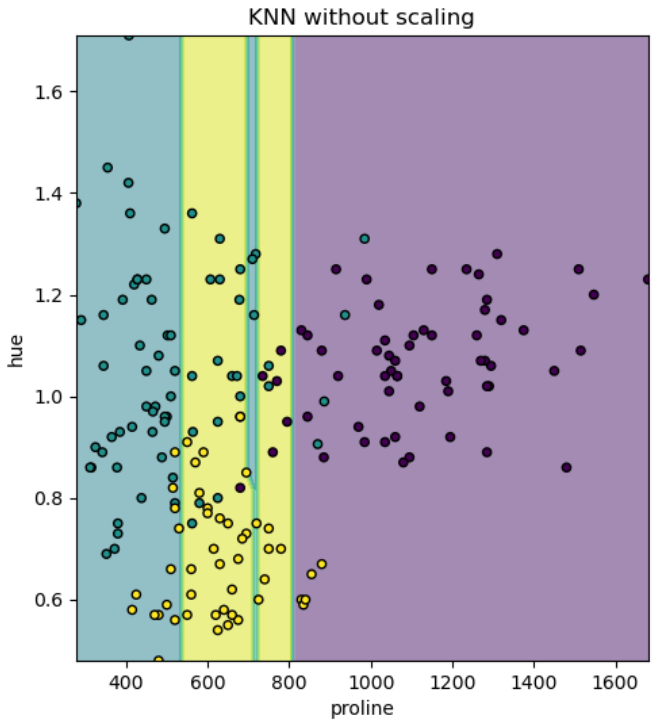


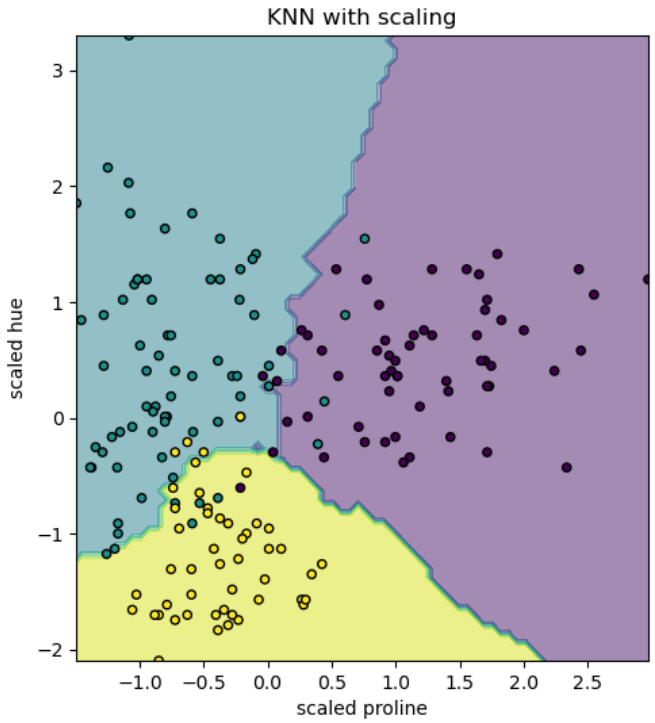
A tie can occur when two or more points are equidistant from an unclassified observation

- Choose different k. For binary classification generally odd numbers are used
- Randomly choose
- Allow observations in until natural stop point

Important Notes : Scaling

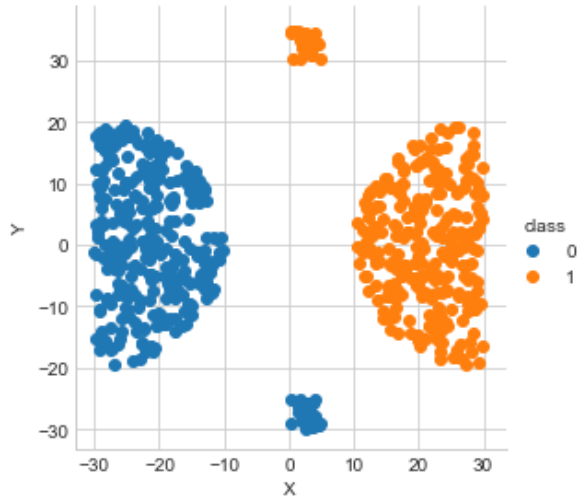


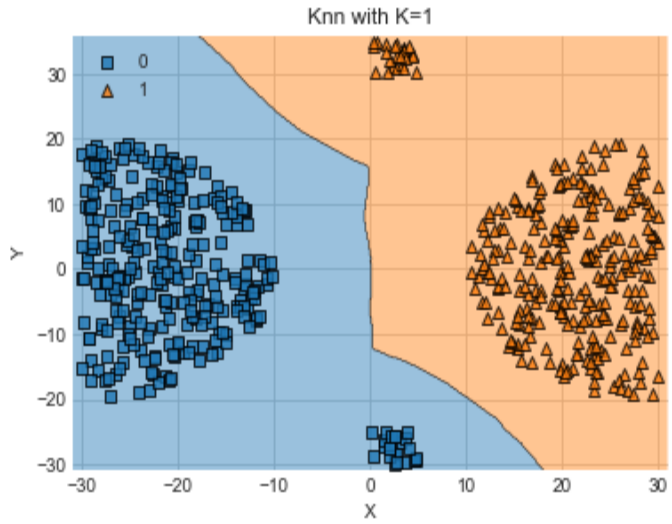


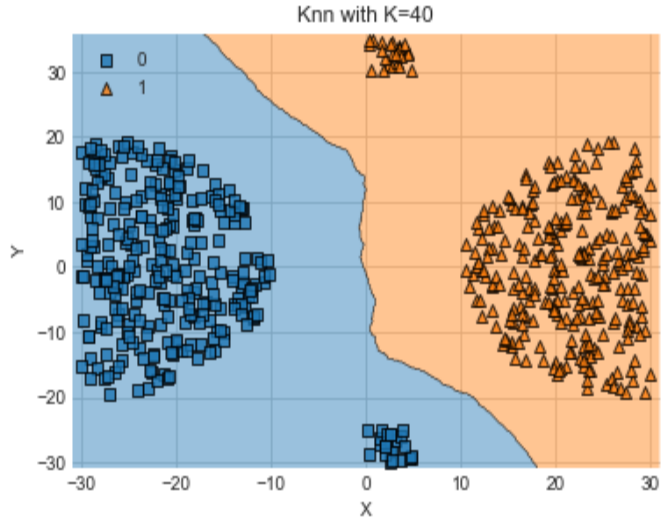


Scaling Methods Future Reading

Important Notes : Outliers







Best Used, Advantages, Disadvantages

Best Used

When you need a simple algorithm to establish benchmark learning rules.

When memory usage of the trained model is a lesser concern.

When prediction speed of the trained model is a lesser concern.

When data set is small.

Advantages

The algorithm is simple and easy to implement.

The algorithm is versatile. It can be used for classification, regression, and search.

There's no need to tune several parameters or make additional assumptions.

Disadvantages

The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

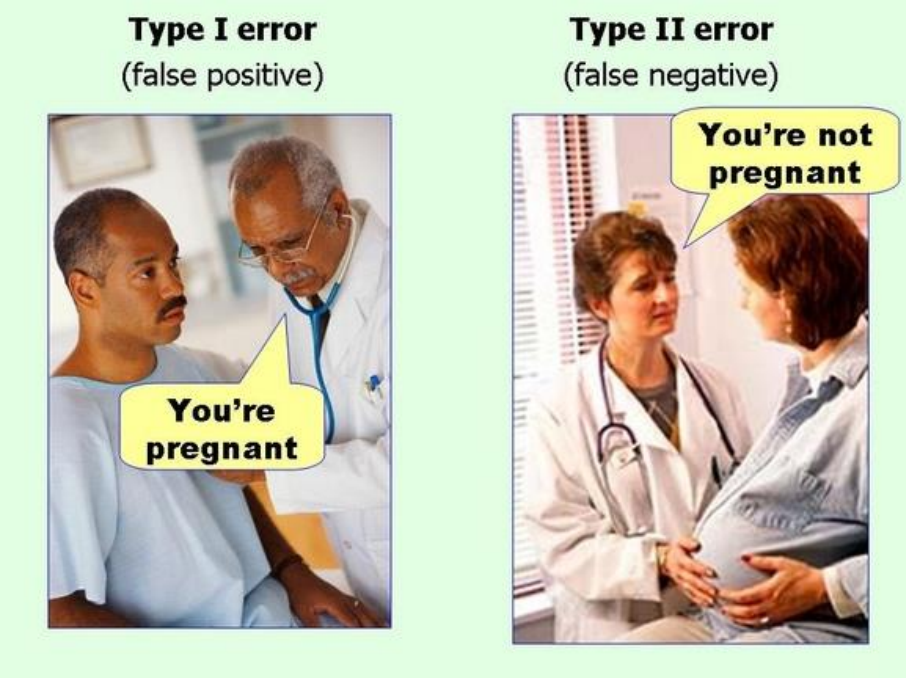
When data set is big.

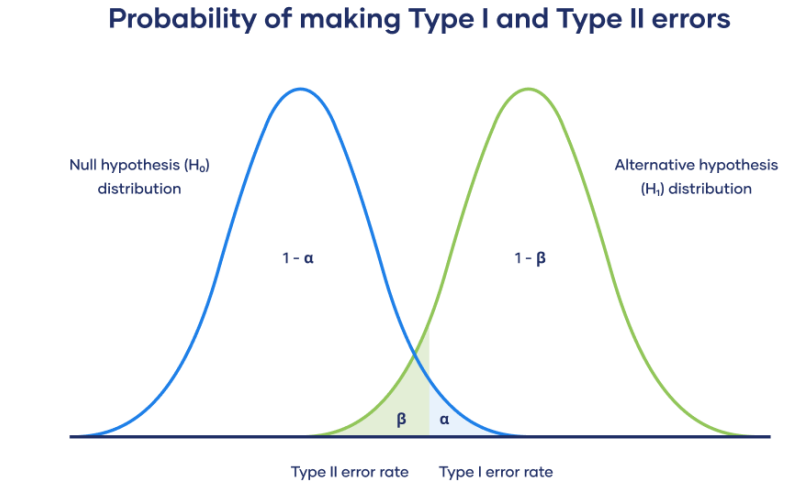
Garanti BBVA Technology

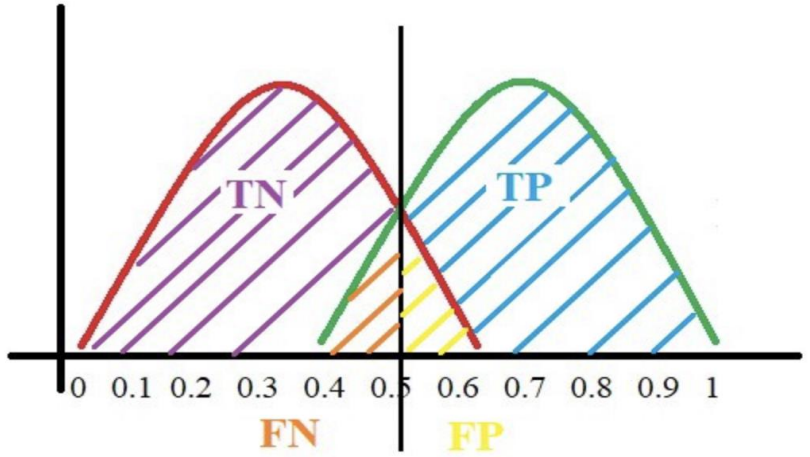
05

Performance Metrics

Performance Metrics : Error Types Revisited







Performance Metrics : Confusion Matrix

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative* Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive* Type I Error	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Why accuracy is not enough?

Future Reading for Confusion Matrix

Performance Metrics : Precision, Recall

Which type of error — FPs or FNs — is more undesirable?

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

	Cancer	No Cancer
Cancer	80	80
No Cancer	20	820

Accuracy : 90%, **Precision:** 20%, **Recall:** %50

- **Precision** measures the extent of **error caused by False Positives (FPs)**
- **Recall** measures the extent of **error caused by False Negatives (FNs)**
- Usually increasing precision will decrease recall, and vice versa.

Future Reading for Imbalanced Datasets

Performance Metrics : F1 Score

What if, the FN and FP errors equally undesirable?

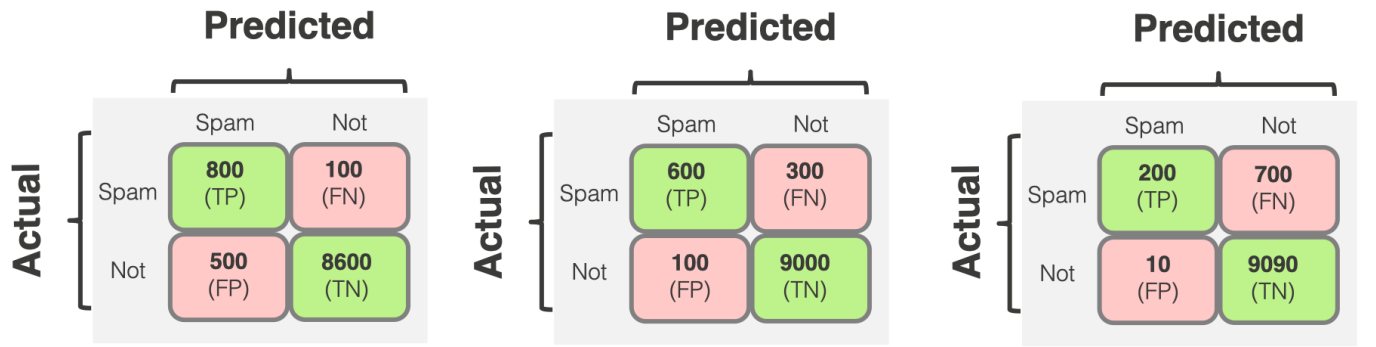
Set 1: (0.5, 0.5)
Set 2: (0.05, 0.95)

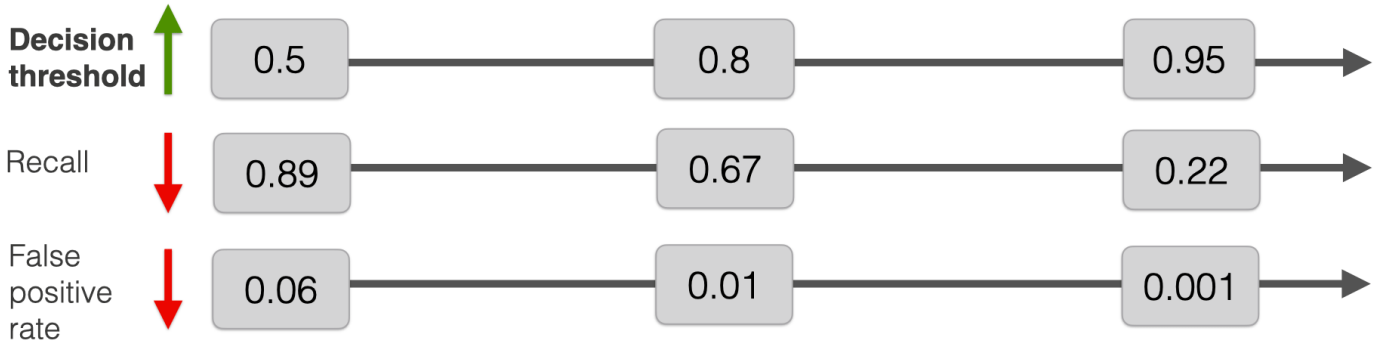
Mean Type	Set 1 Result	Set 2 Result
Arithmetic	0.5	0.5
Geometric	0.5	0.21
Harmonic	0.5	0.09

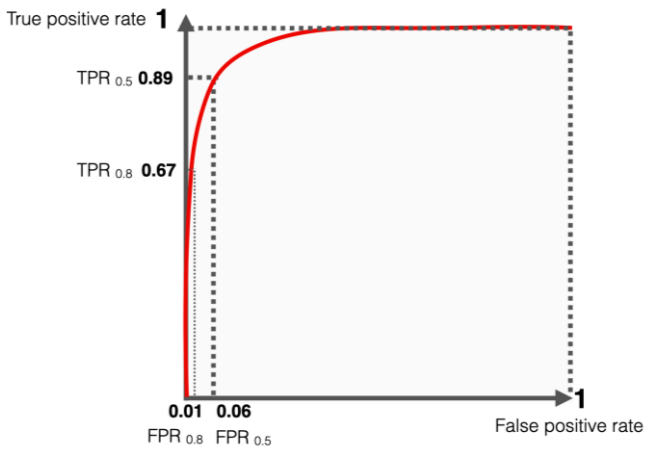
$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

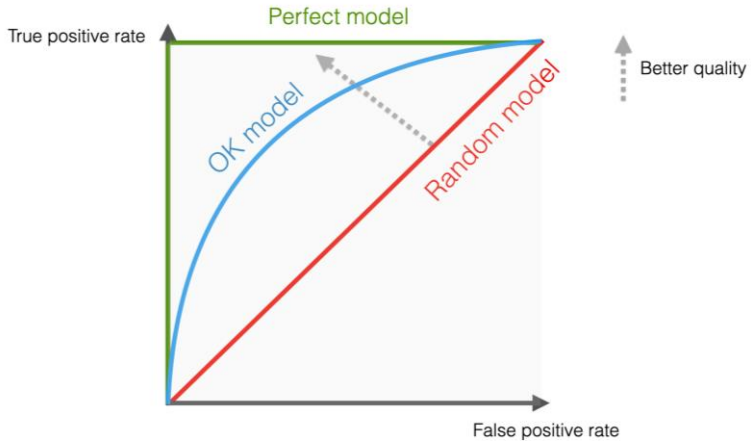
Future Reading for Imbalanced Datasets

Performance Metrics : ROC









$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

Performance Metrics : AUC