Classification

Supervised Machine Learning by tdi@ek.dk

Agenda

- Classification vs Regression
- Algorithms for Classification
  - Decision Tree
  - Random Forest
  - Naïve Bayes
  - Logistic Regression
  - KNN
- Validation of Predictions

What is Classification?

- Machine learning task
- Process of categorising data
- Splitting the data into predefined classes
- Predicting the class of new data

Classification vs Regression

- Both a machine learning approaches for creating a model
- Regression creates a model that best fits our data
- Classification creates a model that best separates the data into classes

Examples

- Predicting which choice the customer would make
- Choosing one of several alternative solutions or outputs
- Recognizing what in on the image or which message is spam spam
- Making medical or technical diagnoses
- Recommending specific products to specific consumers

Classification Task Formulation

If you have a set of objects $X = \{x_1, x_2, \ldots, x_n\}$ and another set of objects $Y = \{y_1, y_2, \ldots, y_m\}$, to which of these two sets you will add a new object Z?

- As a human, you probably base your decision on
  - some rules?
  - proximity?
  - other factors and priorities?

Classification Solution

Human Approach

1. define importance criteria
2. make some rules of choice regarding the importance
3. implement the rules to eliminate some of the choices
4. do this until you end up with a single choice

ML Approach

Same as human approach

- organization of factors, which can influence the decision
- implement the factors to reduce the choices
- end up with one choice

But

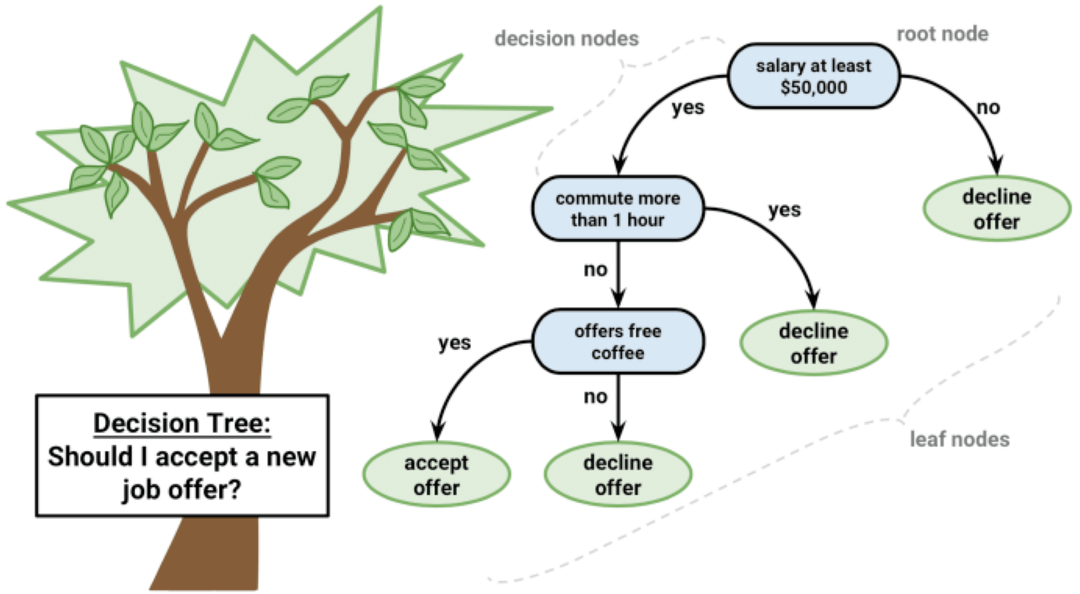
- how to define the rules or the importance criteria?

Decision Tree Algorithms

What is a Decision Tree and how to use it for classification?

Decision Tree

- Hierarchical organisation of decision criteria – like an upside-down tree
- Consists of nodes, branches and leaves
- The top node is called the root node, the nodes under it are child nodes
- All nodes are connected by branches, or edges
- Nodes that are at the end of branches are called leaves

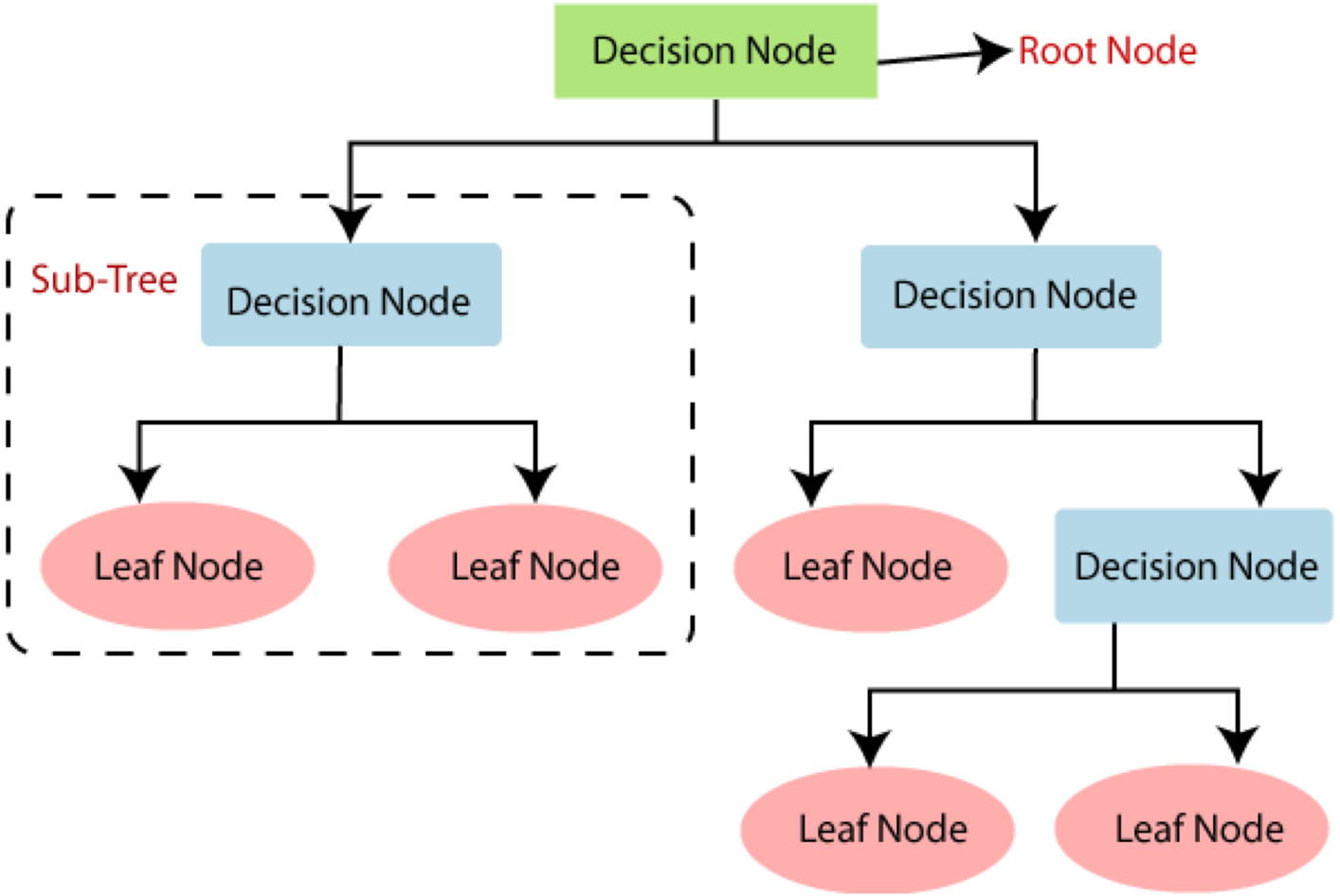


- Each node represents a test on one attribute, and each branch represents a possible outcome of the test
- The leaf nodes of the tree represent the final classifications
- The root node contains the entire data set (all data records), and child nodes hold respective subsets of that set

Nodes and Leaves

Nodes host the rules driving the choices

Leaves host the recommended output



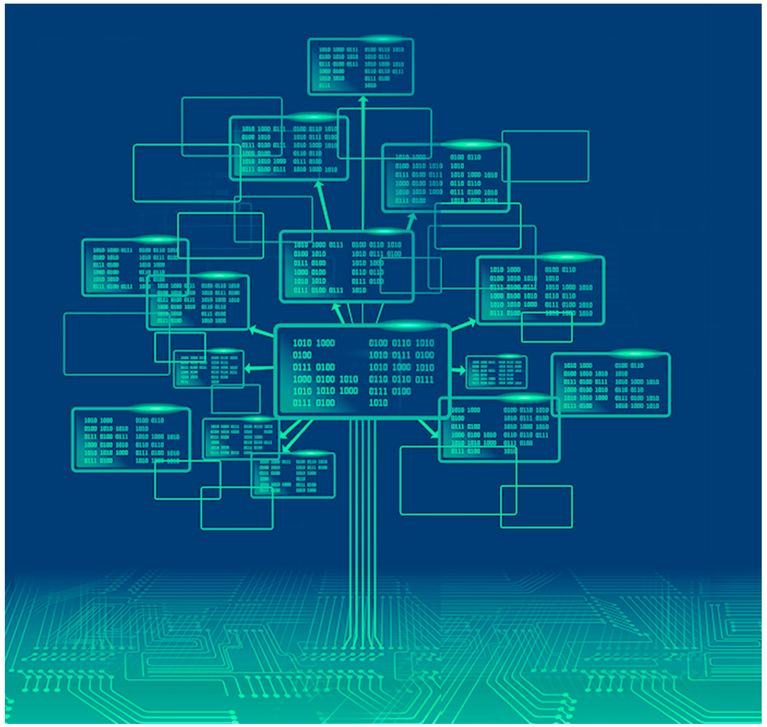
How To Train ML model for Automation?

https://www.youtube.com/watch?v=IpGxLWOIZy4

How To Train ML model for Automation?

https://www.youtube.com/watch?v=IpGxLWOIZy4

Which is the Best Classifier?



- How to choose the first question?
- The one with the maximum separating power
- The one that is most useful in narrowing down the decision space as much as possible
- Eliminates up to half of the options
- A new best is selected at the next level of the hierarchy

https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

How to Measure the Power of Attribute

Metrics from the information theory: Entropy and Information Gain

Entropy is a measure of disorder of a system

In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

(Wikipedia)

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

Entropy

A measure for uncertainty, lack of order or predictability, a degree of randomness

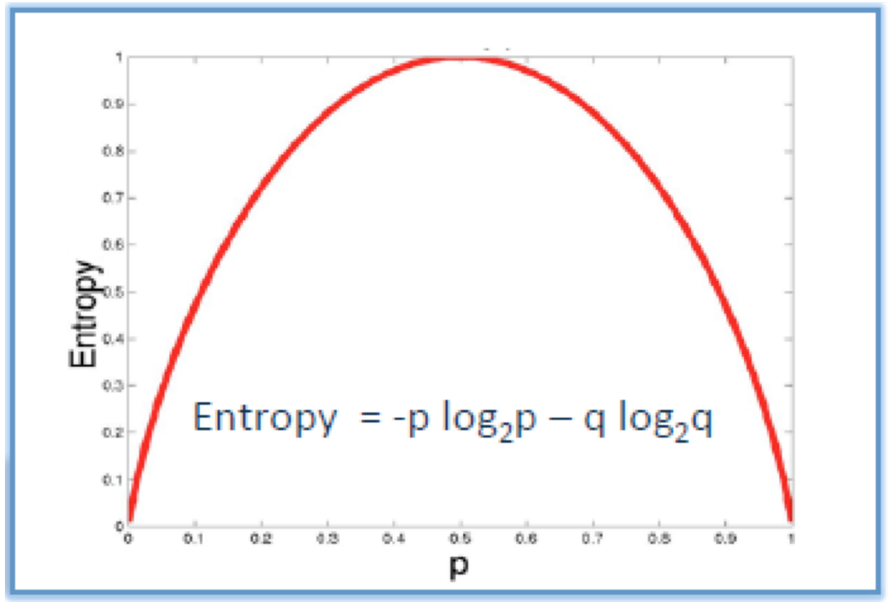
Example:

throw a coin – probability is ½ - max entropy=1

throw a coin, which has identical sides – the result is certain – min entropy

low entropy => order;

high entropy => disorder



Entropy = -0.5 $\log_2 0.5$ – 0.5 $\log_2 0.5$ = 1

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

i – number of classes

p – proportional size of the class

p = class_size / all

Information Gain

In information theory and machine learning, information gain is the amount of information gained about a random variable (X) from observing another random variable (A).

The expected value of the information gain is the mutual information

I(X;A) of X and A

– meaning the entropy of X gets reduced by learning the state of the A

Exercise

Can you predict what I am thinking about?

You can ask me 20 questions for supporting your guess

Try to reduce the entropy and increase the information gain!

Building Decision Tree from Training Data

Greedy: "having or showing an intense and selfish desire for wealth or power"

Definitions from Oxford Languages

- Greedy algorithm - incremental search for the best solution at every step, regardless the following steps.
- ID3 is a classic greedy algorithm used to build a decision tree from the available data as hierarchical structure that can classify data points into different categories.
- ID3 operates with measures, such as entropy and information gain.

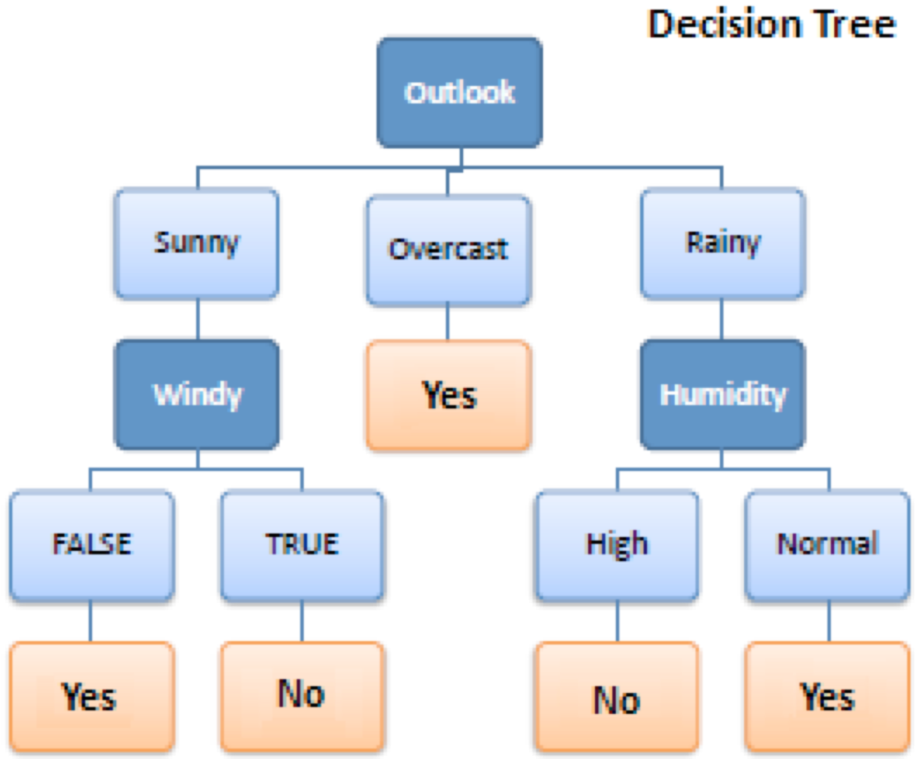
Iterative Dichotomiser 3 (ID3)

- Deals primarily with categorical data
- ID3 recursively splits the data set into smaller and smaller subsets until reaching a subset, in which all data points belong to the same class
- The split starts with that attribute, which provides the most information about the dependant variable – it will be at the root of the tree
- this attribute best separates the data points into different subsets – reduces the entropy, providing most information gain
  - for each of the subsets, the algorithm searches again for another attribute, which is the best in that subset, and so on.
  - the best attribute in one subset is the one with lowest entropy and highest information gain – the algorithm calculates these at every node

Example: Planning a Game of Golf Depending on the Weather Attributes

Predictors | Target

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



ID3 Decision Tree Algorithm

Step 1: Calculate entropy of the parent

Step 2: Split it and then calculate the weighted sum of the entropy of the children

- the total entropy of children is subtracted from the entropy of the parent
- the result is the Information Gain, or the decrease in entropy

Step 3: For a decision node, choose the candidate attribute with the max information gain

- divide the dataset by the branches of this attribute
- repeat the same process on every branch

Step 4a: A branch with entropy of 0 (zero) is a leaf node, and is used for making predictions

Step 4b: A branch with entropy more than 0 needs further splitting

Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified

Decision Tree Implementation

Once created, the decision tree can be navigated and a path from the root to one leave to be found for every new row of input data.

Decision Tree can also be transformed into set of rules by mapping from the root node to the leaf node

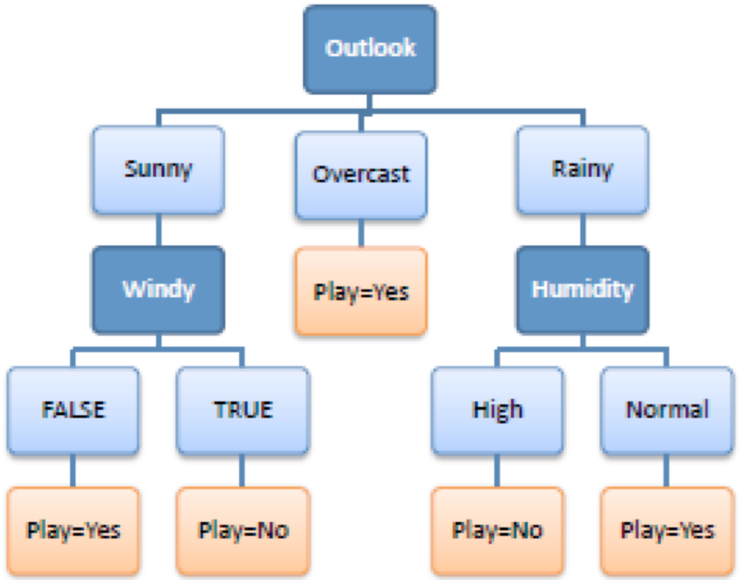
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

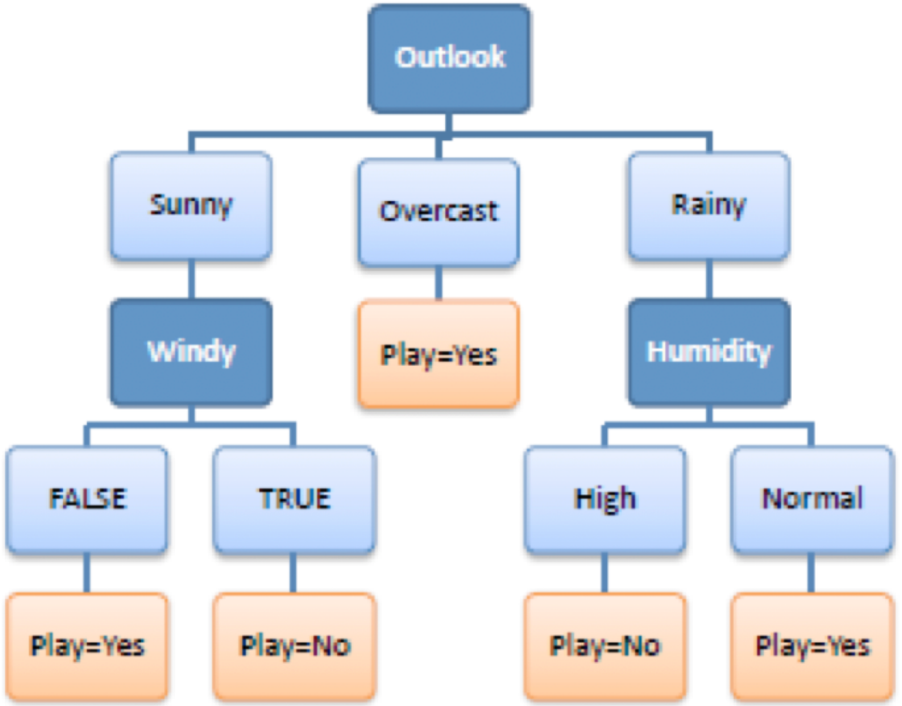
R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



The rules are applied, when a new case needs to be classified.

Exercise:
Apply this Decision Tree to predict the output?

Overlook	Temperature	Humidity	Wind	Will I play golf?
sunny	hot	normal	false	
sunny	cold	normal	true	
rainy	cold	high	true	



Advantages and Disadvantages of Decision Trees

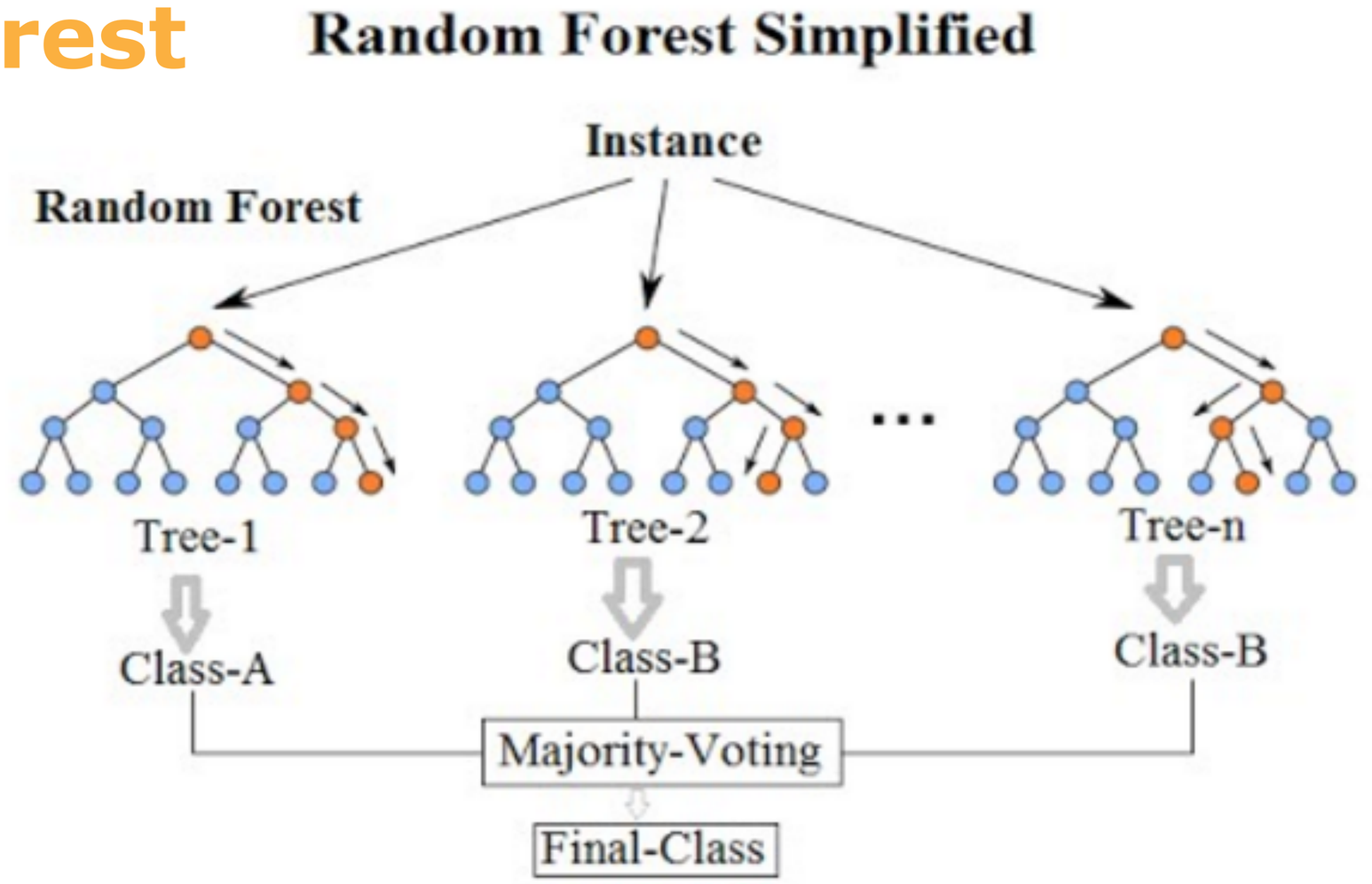
Advantages

- Visualises the solution
- Easy to follow any path through the tree
- Relationships discovered by a decision tree can be expressed as a set of rules

Disadvantages

- Not suitable for continuous data
- Can produce reliable outcomes only from clean data
- Inability to examine more than one variable at a time

Random Forest



Naïve Bayes

Based on Probability

Thomas Bayes
statistician
philosopher
priest

Thomas Bayes
(1701 - 1761)

He is known to have published two works in his lifetime, one theological and one mathematical:

1. *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* (1731)
2. *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst* (published anonymously in 1736), in which he defended the logical foundation of Isaac Newton's calculus ("fluxions") against the criticism by George Berkeley, a bishop and noted philosopher, the author of *The Analyst*

Bayes' Theorem

- A and B are two different events
- P(A): probability that A happens (prior probability)
- P(B): probability that B happens (prior probability)
- P(A|B): probability of A happens, given B happens
- P(B|A): probability of B happens, given A is happens
- P(B) ≠ 0

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Example

- David and John are invited to a party
- What is the probability that both of them will accept the invitation?

Probability Theory

John says: if David goes, I may go, too

Let

P(D) - the probability David is going

P(J) - the probability John is going

David says: if John goes, I may go, too

P(J|D) – if David goes to a party, John goes with him

P(D|J) – if John goes to a party, David goes with him

Naïve Bayes Classification Algorithm

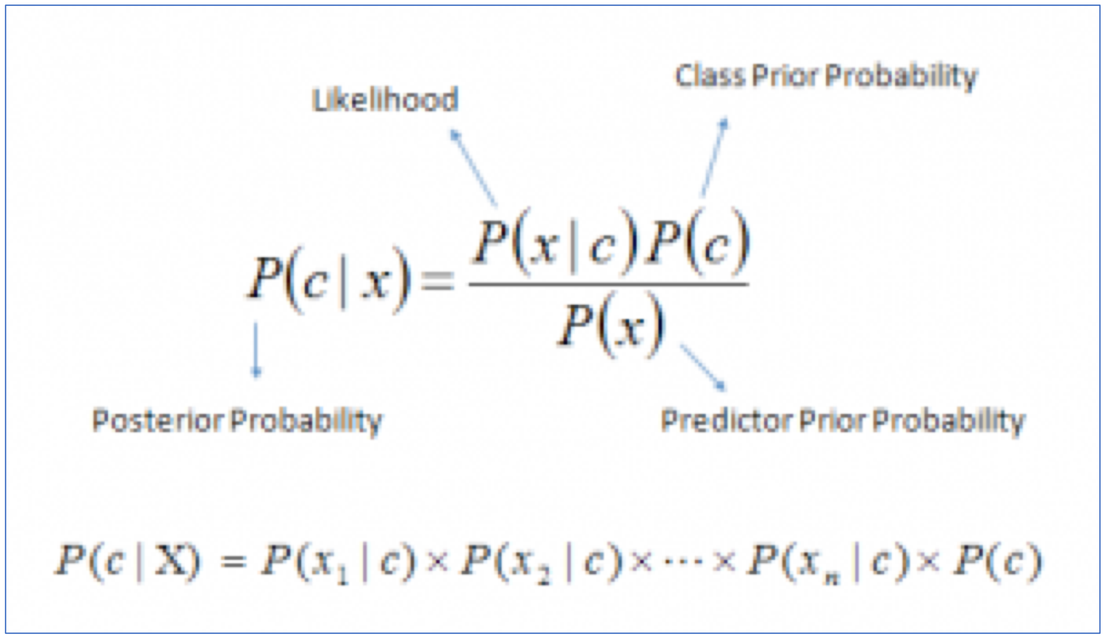
C1, C2 – optional output classes

$X=(x_1, x_2, x_3, \ldots x_n)$ – predictor

P (C|X) – posterior probability

There are several optional output classes

- calculate the probability of each
- select the most probable one



Example:
Who is the author?

Who has written these love letters?

"I love the winter season. During my ski holiday I had wonderful experience and met great people. "

I am writing to tell you how wonderful you are and how much I love you.

Alice

Bob

Bayes Advantages and Disadvantages

- Easy and fast to predict class of test data set
- Performs well in multi class prediction
  - performs better compared to other models like logistic regression
  - you need less training data
- Performs well in case of categorical input variables compared to numerical input
  - for numerical variables, normal distribution is assumed (a strong assumption) – Gaussian

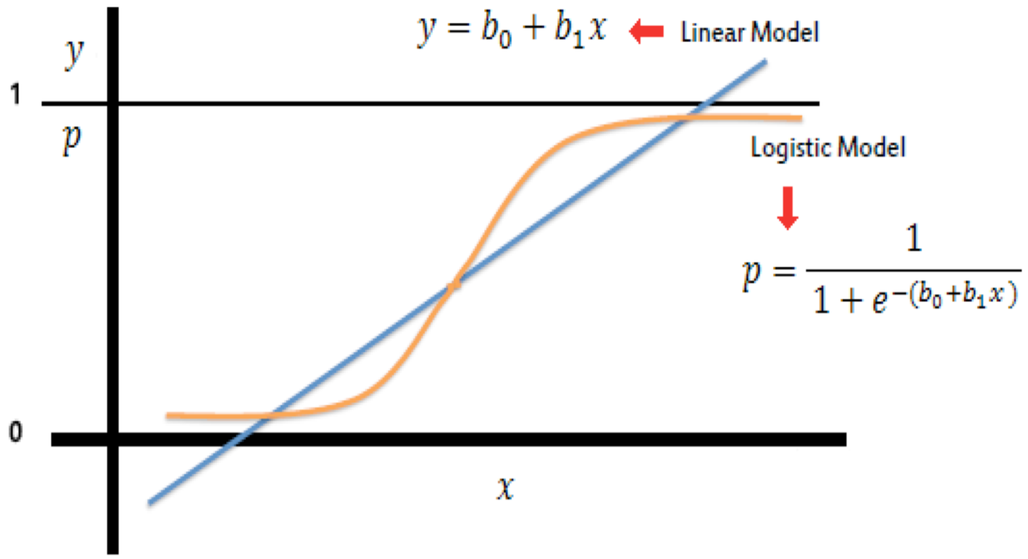
- Limitation of Naïve Bayes is the assumption of independent predictors (naïve)
- If categorical variable has a value not observed in training data and seen in test data set, the model will assign a 0 (zero) probability and will be unable to make a prediction
- Naïve Bayes is known as a bad estimator with no precise results

Logistic Regression

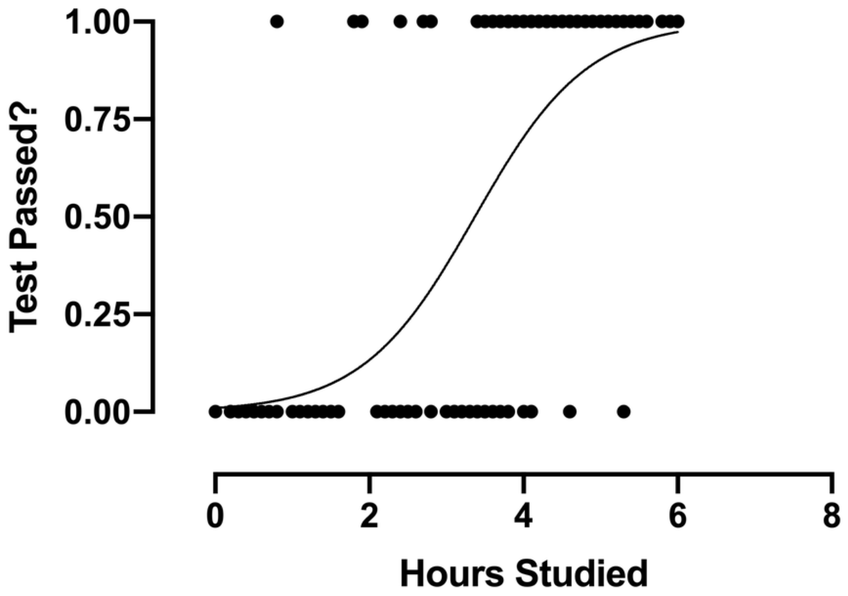
Compared to Linear Regression and Bayesian

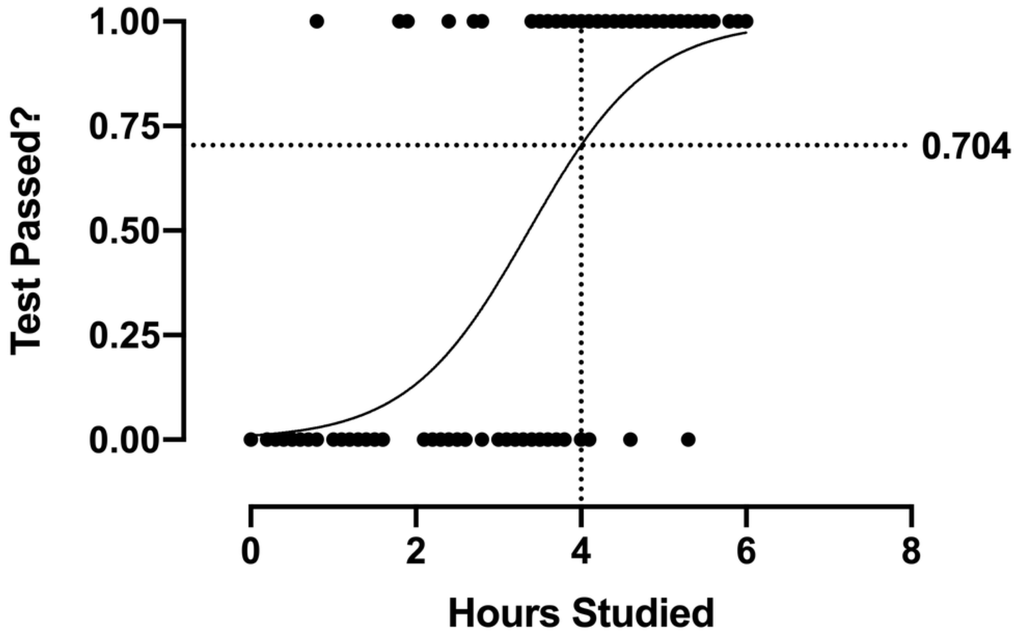
Logistic Regression

- Statistical method for predicting a categorical outcome by modeling the probability of an event
  - like yes/no, approved/rejected
- Predicts the probability of a binary outcome - 0 or 1
  - probability is always between 0 and 1
- Using S-shaped logistic function
  - to keep y value in [0, 1]
  - uses threshold to classify the final result



Example





Threshold

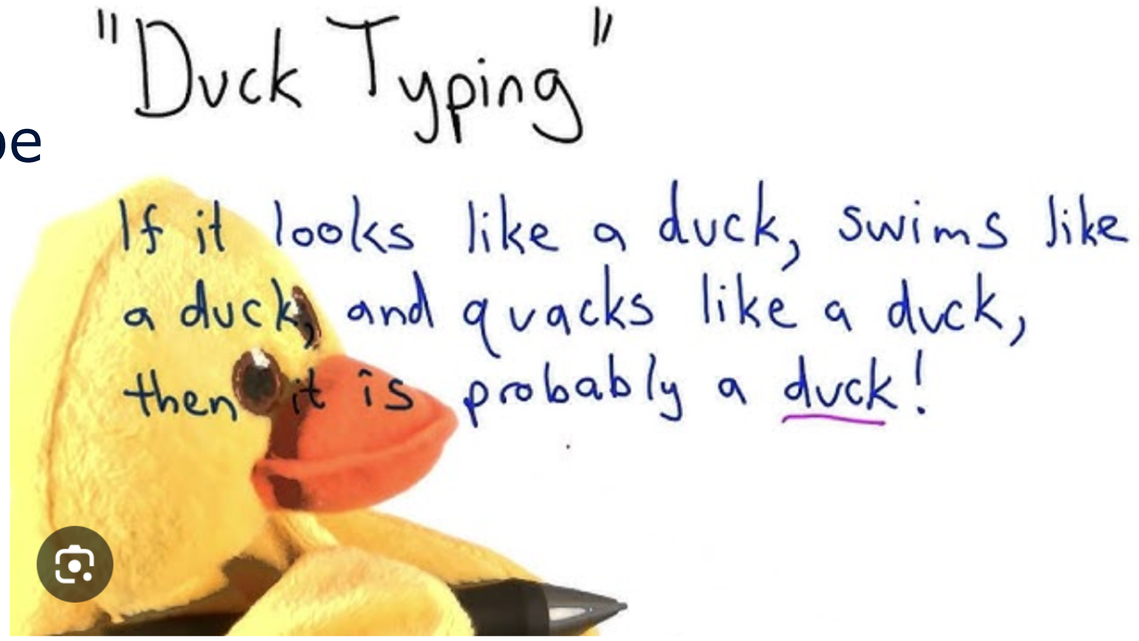
In 70% of cases with 4 hours study the prediction “pass ” will be correct

KNN

K Nearest Neighbours

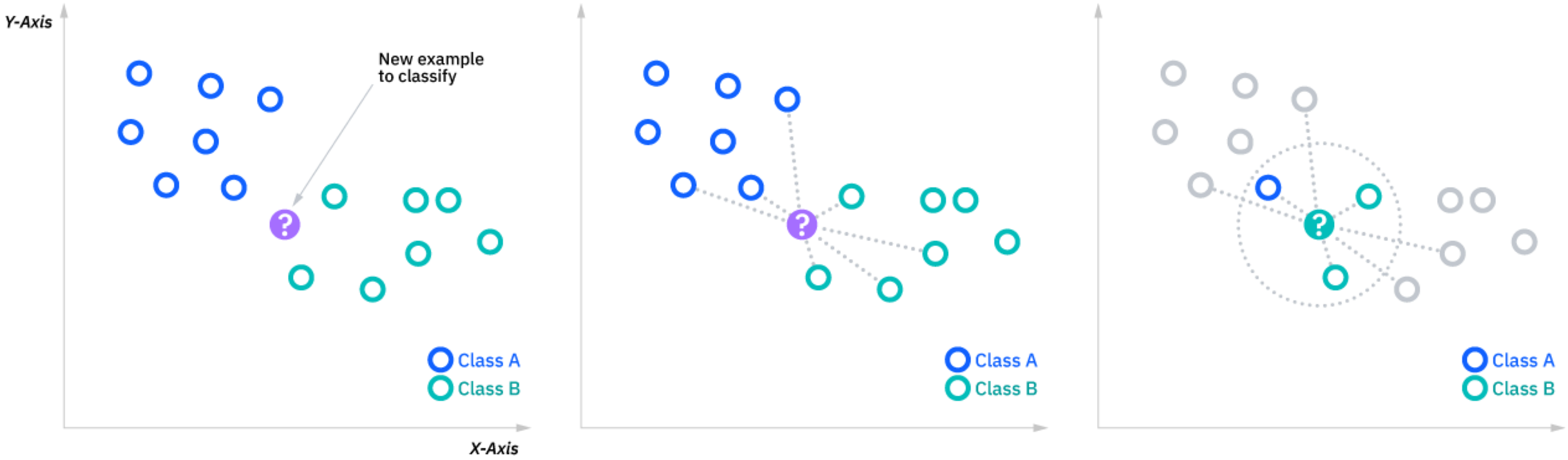
KNN Algorithm

- Supervised learning classifier
  - uses proximity and voting methods to make classifications or predictions
  - assumption that similar points can be found near one another
  - can be used for either regression or classification
  - one of the popular and simplest classification and regression algorithms, mostly used for classification



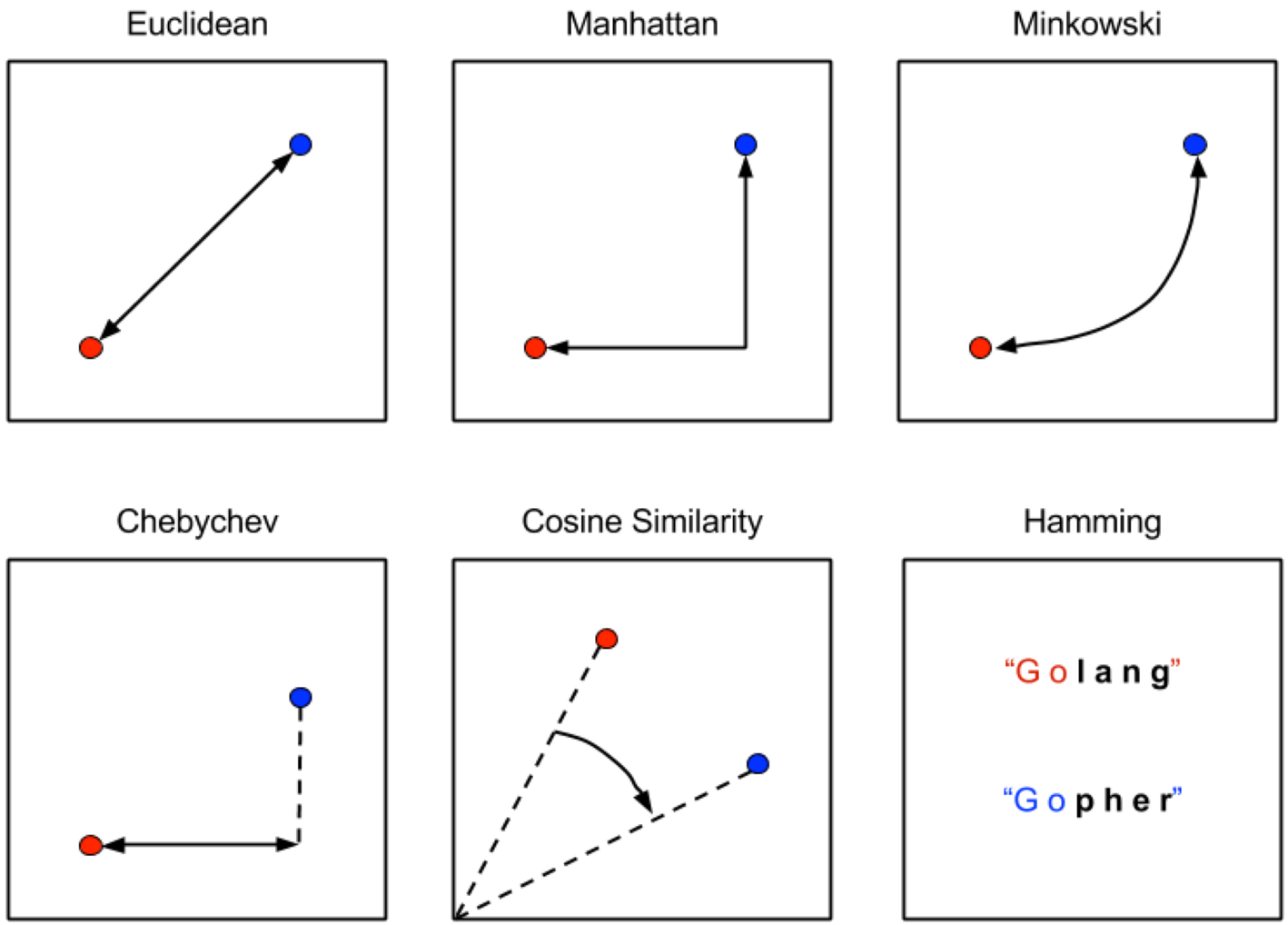
https://www.youtube.com/watch?v=WM1x2L1dFJk

KNN Algorithm



https://www.ibm.com/think/topics/knn

Popular Distance Measures in ML



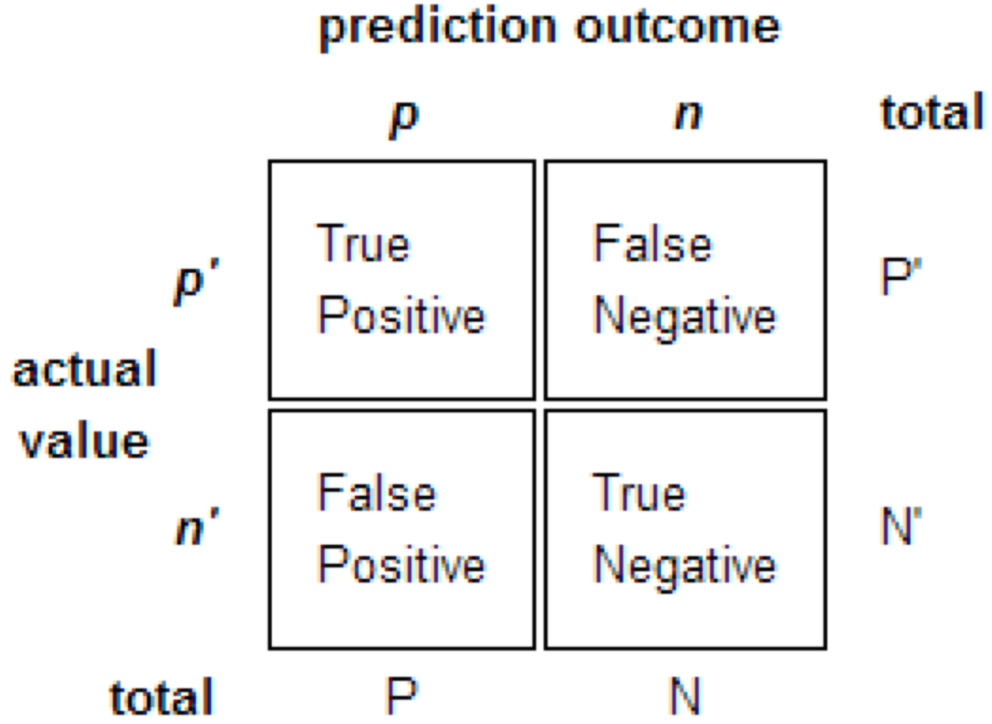
https://tuhinmukherjee74.medium.com/different-types-of-distances-used-in-machine-learning-explained-550e2979752c

Validation of Predictions

Measures for Model Quality

Confusion Matrix

- Searching for estimation of the difference between the predictions and the observations
- Based on number of cases of correct or wrong prediction



Confusion Matrix

Validation measures

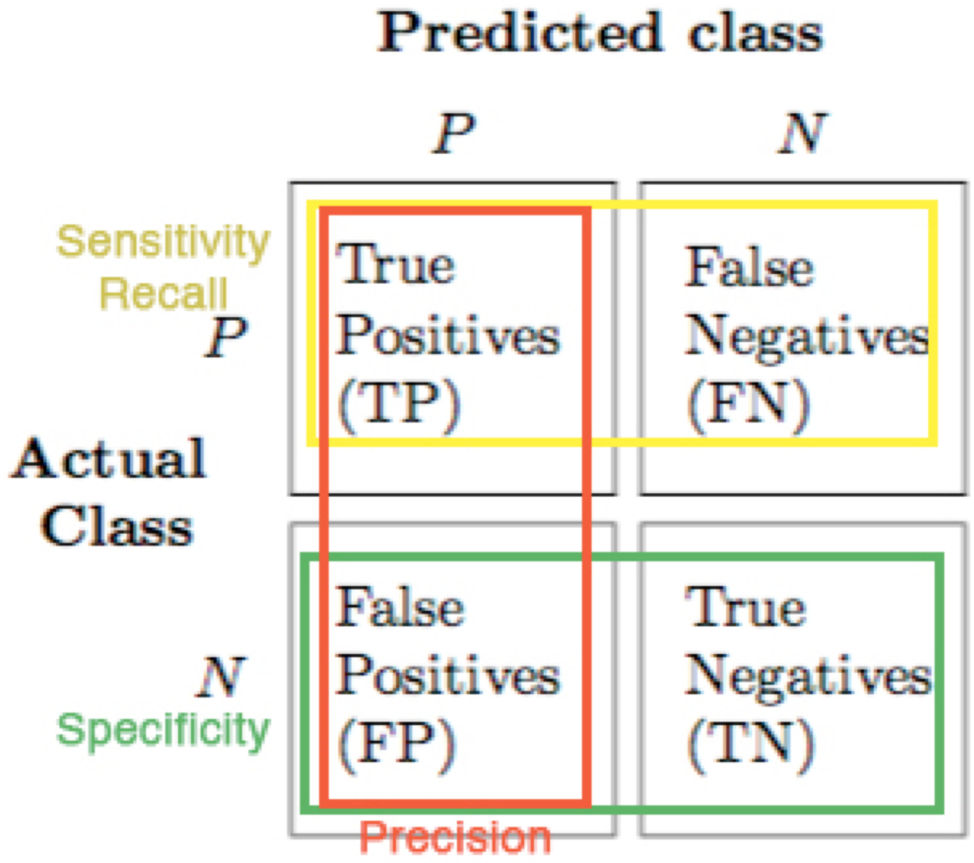
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1 = (2*Precision*Recall)/(Precision + Recall)
- F1 is Harmonic Mean of Precision and Recall

If a and b are postive numbers, then

$$\text{Arithmetic Mean (AM)} = \frac{a+b}{2}$$

$$\text{Geometric Mean (GM)} = \sqrt{ab}$$

$$\text{Harmonic Mean (HM)} = \frac{2ab}{a+b} = \frac{(GM)^2}{AM}$$



Reference

- Prateek Joshi, Artificial Intelligence with Python
- Russell & Norvig, AI: A Modern Approach, 3rd Ed.
- http://www.saedsayad.com/decision_tree.htm
- https://www.youtube.com/watch?v=IpGxLWOIZy4
- http://www.cedar.buffalo.edu/~srihari/CSE574/Chap16/Decision%20Trees.Part1.pdf
- https://towardsdatascience.com/entropy-the-pillar-of-both-thermodynamics-and-information-theory-138d6e4872fa
- http://archive.ics.uci.edu/ml/index.php
- https://www.youtube.com/watch?v=IpGxLWOIZy4:5:58
- https://justinmaes.wordpress.com/2016/10/08/intro-to-machine-learning-naive-bayes-part-2/
- https://appliedmachinelearning.blog/2017/05/23/understanding-naive-bayes-classifier-from-scratch-python-code
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php