GATE by moyyn

Data Analyst Module 4

Dr. Kavitha Chetana Didugu

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law
Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

STARTUP SCHOOL by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin
Der Regierende Bürgermeister von Berlin
Senatskanzlei
Wissenschaft und Forschung

Today’s Agenda

Back to Basics

- Predictive Analytics: Introduction
- Models and Metrics
- Code Example

Predictive Analytics using ML

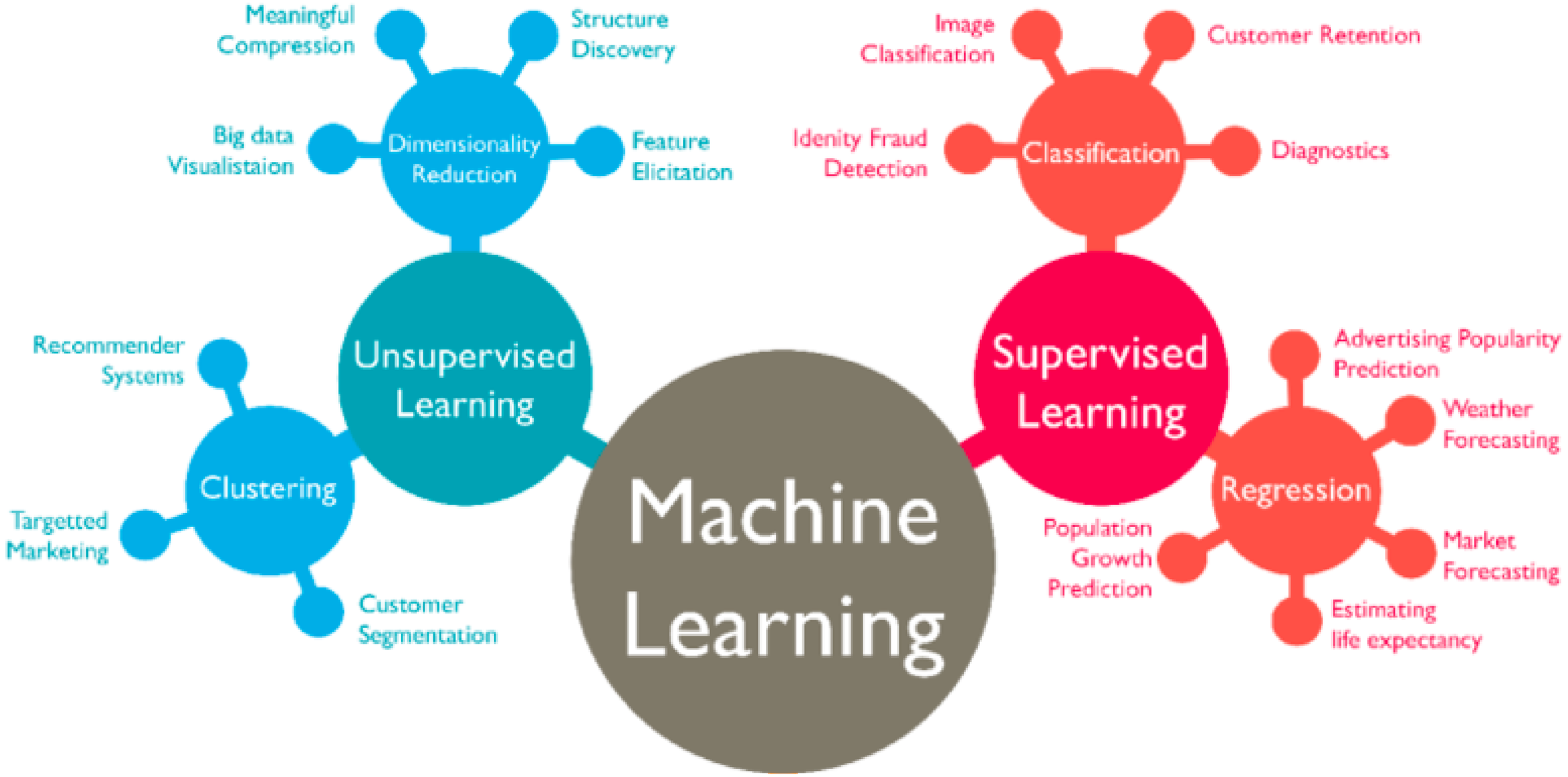
Machine Learning is no longer a Data Scientist's niche!

What is Predictive Analytics

Predictive Analytics is the use of Data to identify patterns, predict trends, etc.

Types of Predictive Analytical Tasks

- Regression: price prediction, income predictions, etc.
- Time Series Forecasting: predicting future trends of - sales, stock price, etc.
- Classification: predicting whether it will rain tomorrow or not (weather prediction), predicting whether a loan applicant’s application will be accepted or rejected, etc.
- Anomaly Detection:
- Clustering: Finding out similarities within data and grouping them into similar clusters- customer segmentation, topic modelling, etc.
- Recommendation Systems: based on user history, suggest the right product



Taxonomy of ML

Objective Function and Performance Metrics

The most important parts of an ML model

What is an Objective Function?

Objective Function is a Mathematical Representation of the Goal that the model needs to achieve to fit the model on the data in the best possible way.

An Objective Function always needs to be optimised: MINIMUM POSSIBLE ERROR, MAXIMUM POSSIBLE ACCURACY

Some Examples

- Regression: (linear regression) Ordinary Least Square- least square error
- Classification: Maximum Likelihood Estimate
- Clustering: Total distance of each datapoint from its respective cluster centre

What is a Performance Metric?

A measure of how far the model is from its goal

Some Examples

- Regression: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-square (Amount of variance in the data that is captured by the model)
- Classification: Probability of misclassification (binary cross entropy, categorical cross entropy), Purity of classes within each resultant group, Accuracy, Precision, Recall, F1, AUC, ROC
- Clustering: Davies-Bouldin Index (ratio of within cluster distances to between cluster distances)

Supervised Learning: Regression

Regression

Supervised Learning Model: given a set out inputs, we know exactly what the input is. The model is supervised by the outputs in reaching the best pattern. Regression uses input features/variables based on how strongly they correlate to the output variable

What straight line/linear function fits the data best? Which model explains the variation in the data the best?

How do we know what is BEST: high accuracy? least error?

A sample problem: https://tinyurl.com/kaggleds1

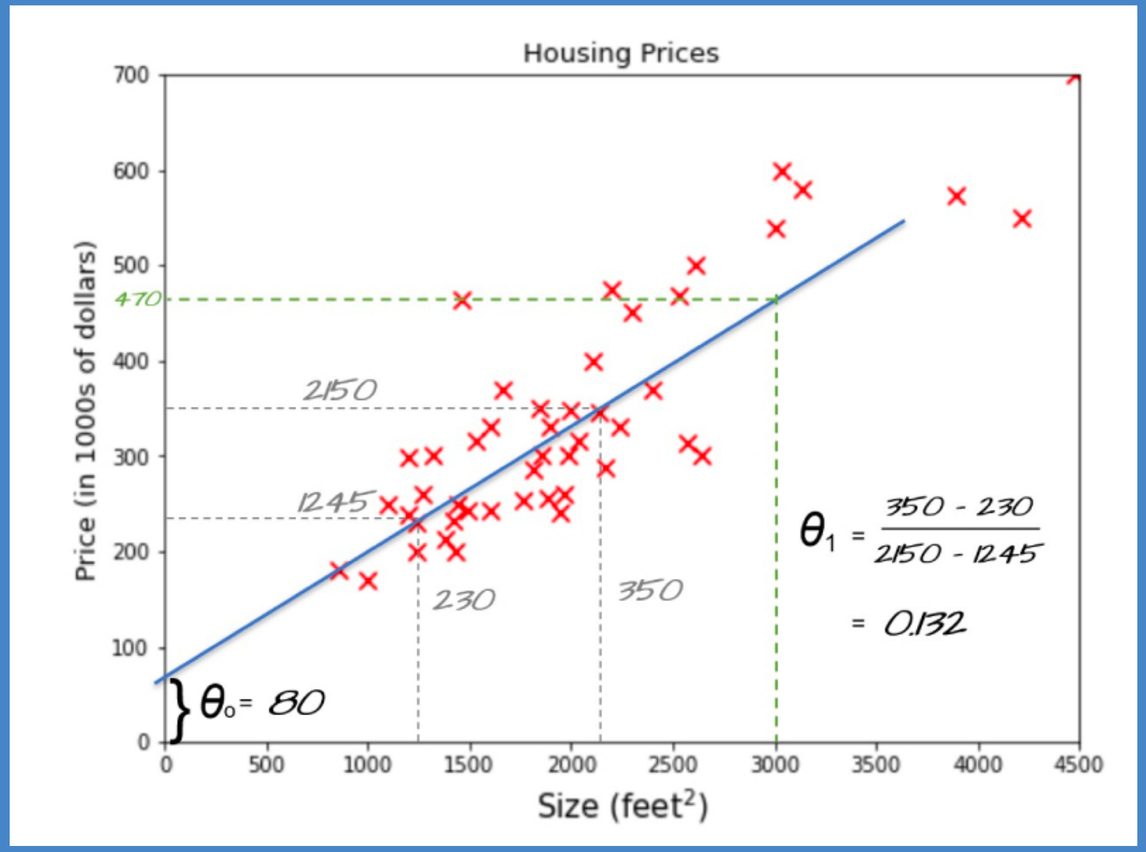


Image reference: https://tinyurl.com/regressionimage

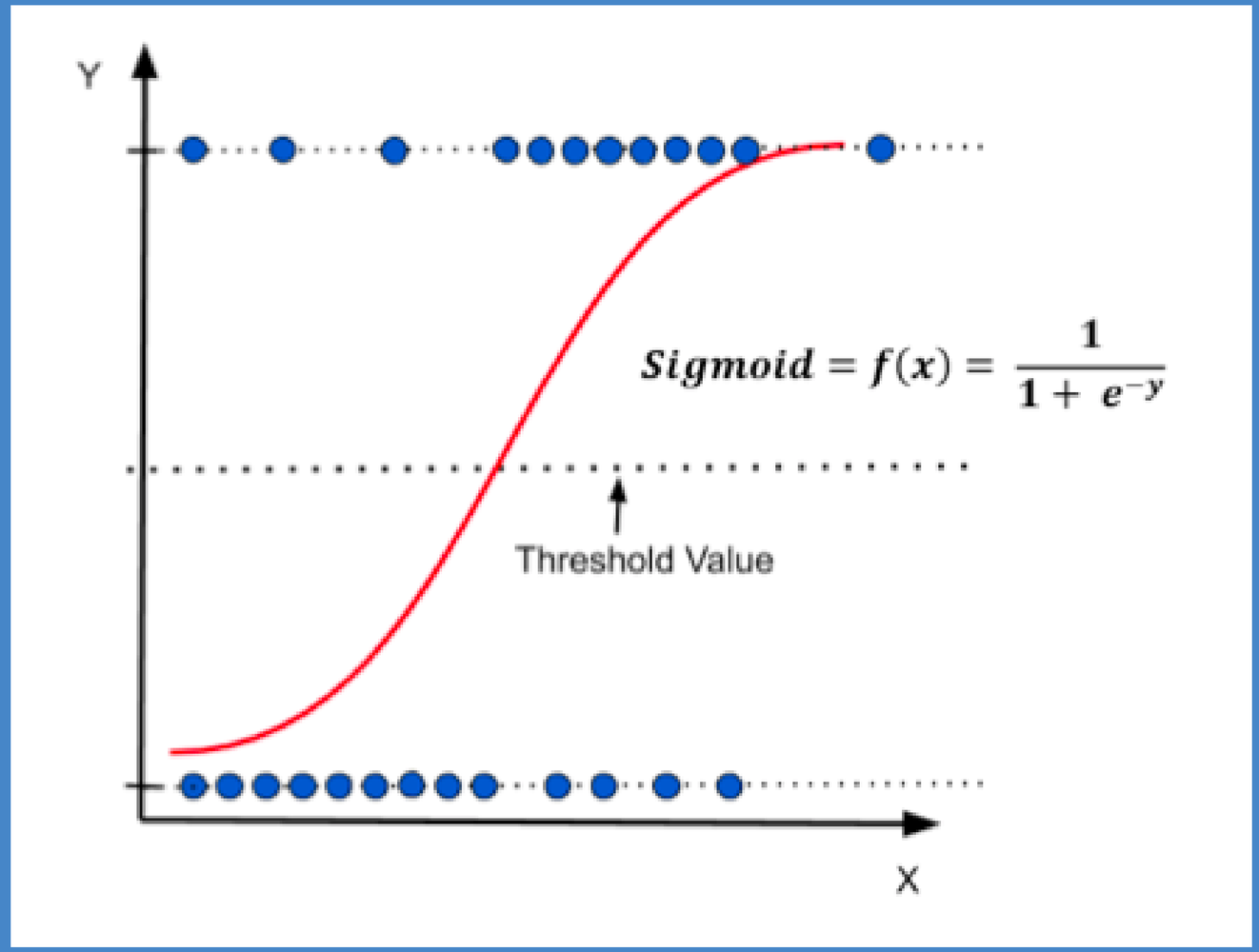
Supervised Learning: Classification

Definitely not just some lines of Code!!

Logistic Regression

Though it is named Regression, it is a Classification model

- Works when there are only two possible outcomes: binary classification
- Predicts probability that a particular data point belongs to a certain class
- the model must be able to predict all probabilities for all records it is being trained on: product of multiple probabilities- very complex model
- Simplified: Sigmoid function- pushes high positive values to probability 1 (class 1) and high negative values to probability 0 (class 2), 0 value corresponds to probability 0.5



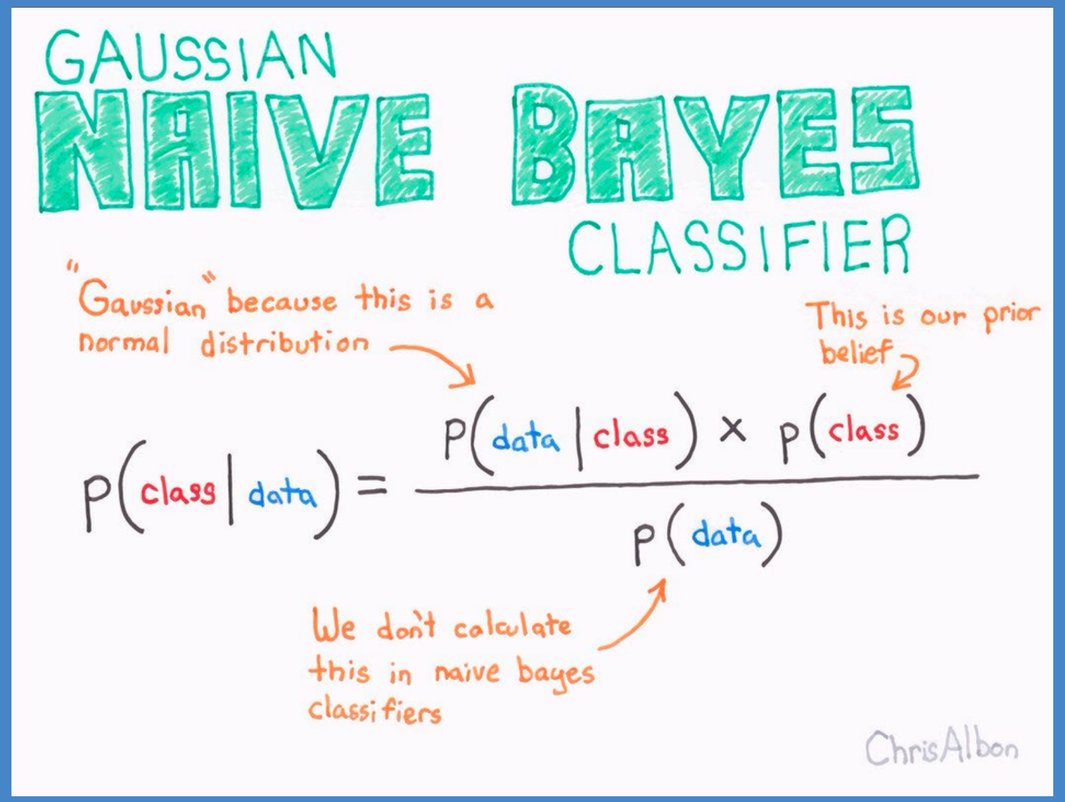
https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2F%40MudSnail%2Fthe-importance-of-logistic-regression-in-image-classification-1966d07e7a0c&psig=AOvVaw1n1BH7yLdMYjM3TAOp6PhV&ust=1697004559577000&source=images&cd=vfe&ved=0CBIQjhxqFwoTCKi4wOHo6oEDFQAAAAAdAAAAABAE

What happens when we have more than 2 classes?

We need more flexible models

NOTE: All models we discuss from now on, have a Classifier variant and a Regressor variant.

Naive Bayes Classifier



- Based on Bayes Theorem of Conditional Probability
- Model tries to maximise the Probability that given the data (all observations in the dataset), what is the probability that each datapoint belongs to a particular class

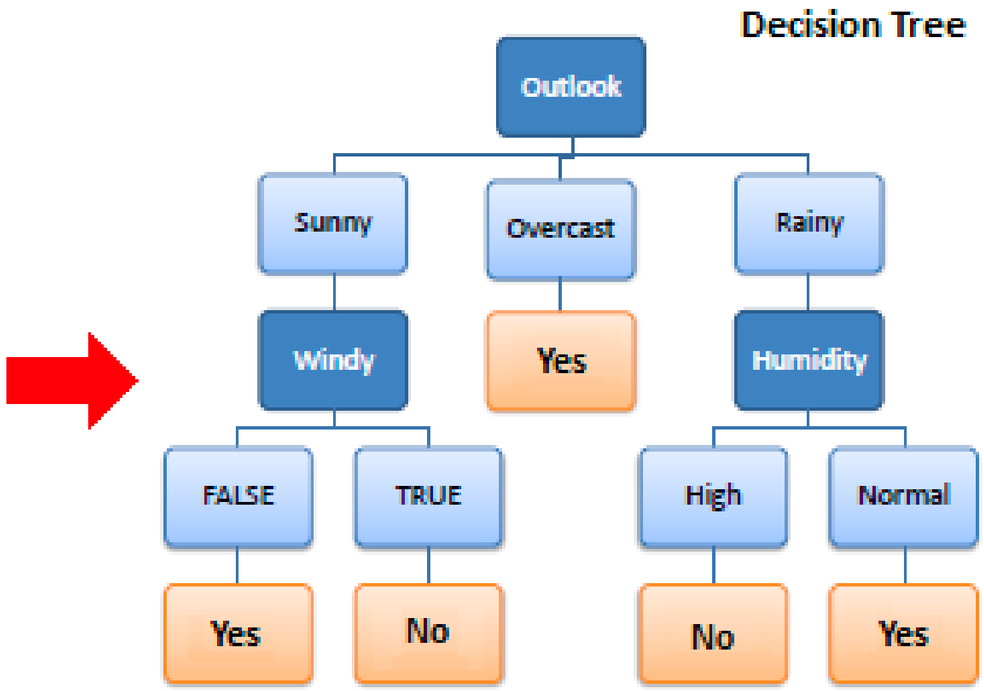
NOTE: *(Gaussian) Naive Bayes assumes that each feature follows a Normal Distribution, and that each input feature is independent from the other.*

Decision Tree

Not much math, just a lot of logic!

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

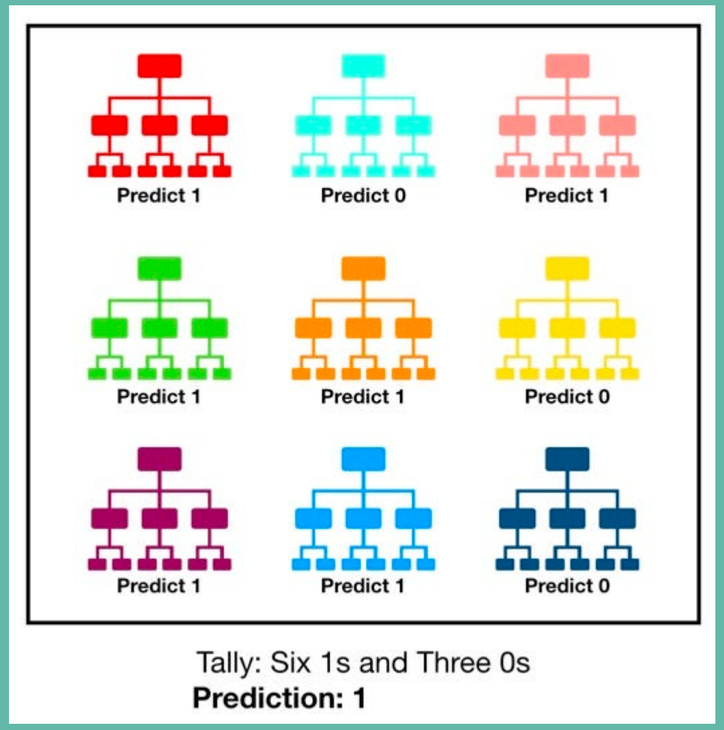


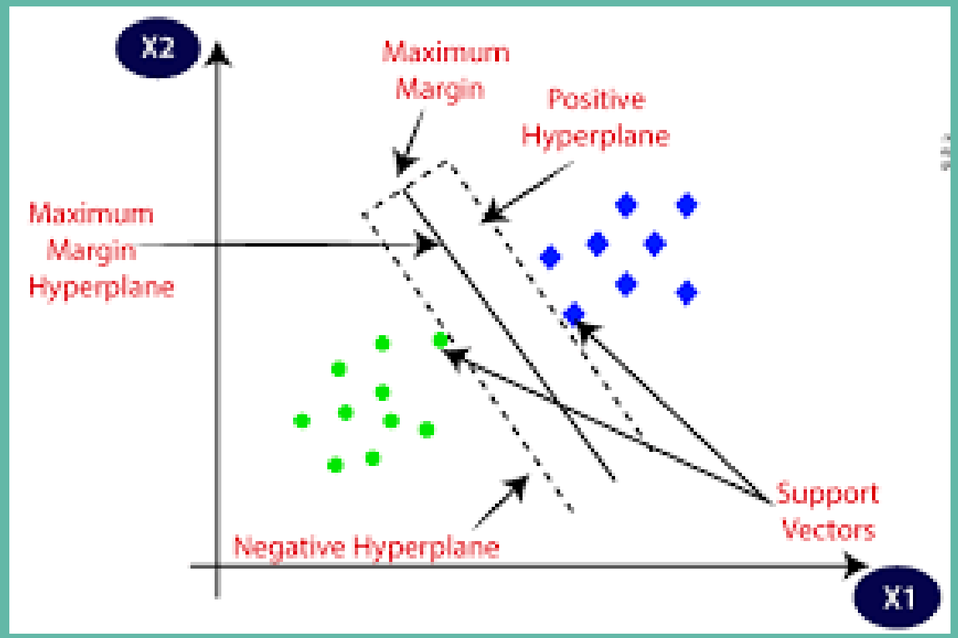
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.saedsayad.com%2Fdecision_tree.htm&psig=AOvVaw0fPLNuI2PpU_cXKodzDYCR&ust=1697629601029000&source=images&cd=vfe&opi=89978449&ved=0CBQQ3YkBahcKEwjAgvWcgf2BAxUAAAAAHQAAAAQBw

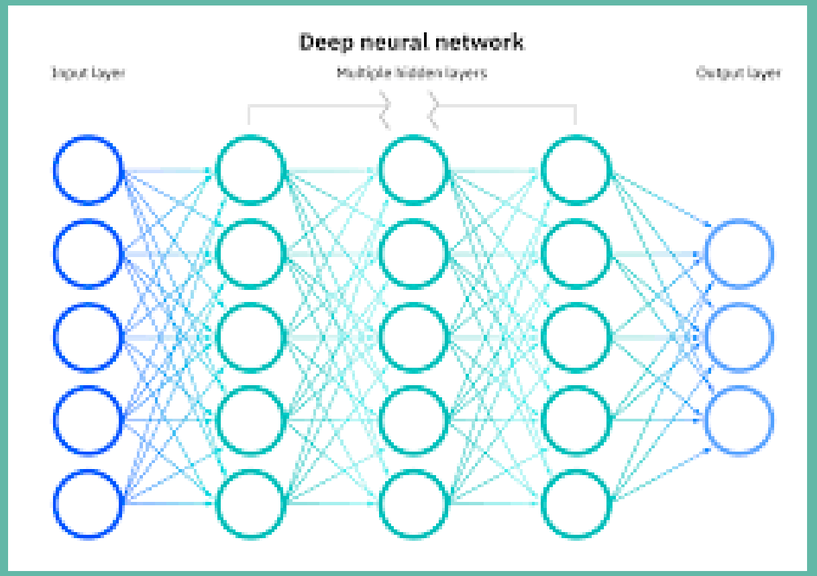
But DTrees are Weak learners, prone to overfit

Other Popular Models

- Random Forest
- XGBoost
- LightGBM
- Support Vector Machines (SVMs)
- Neural Networks







Unsupervised Learning: Similarity and Clustering

Definitely not just some lines of Code!!

Some Common Use-cases

Fraud Detection

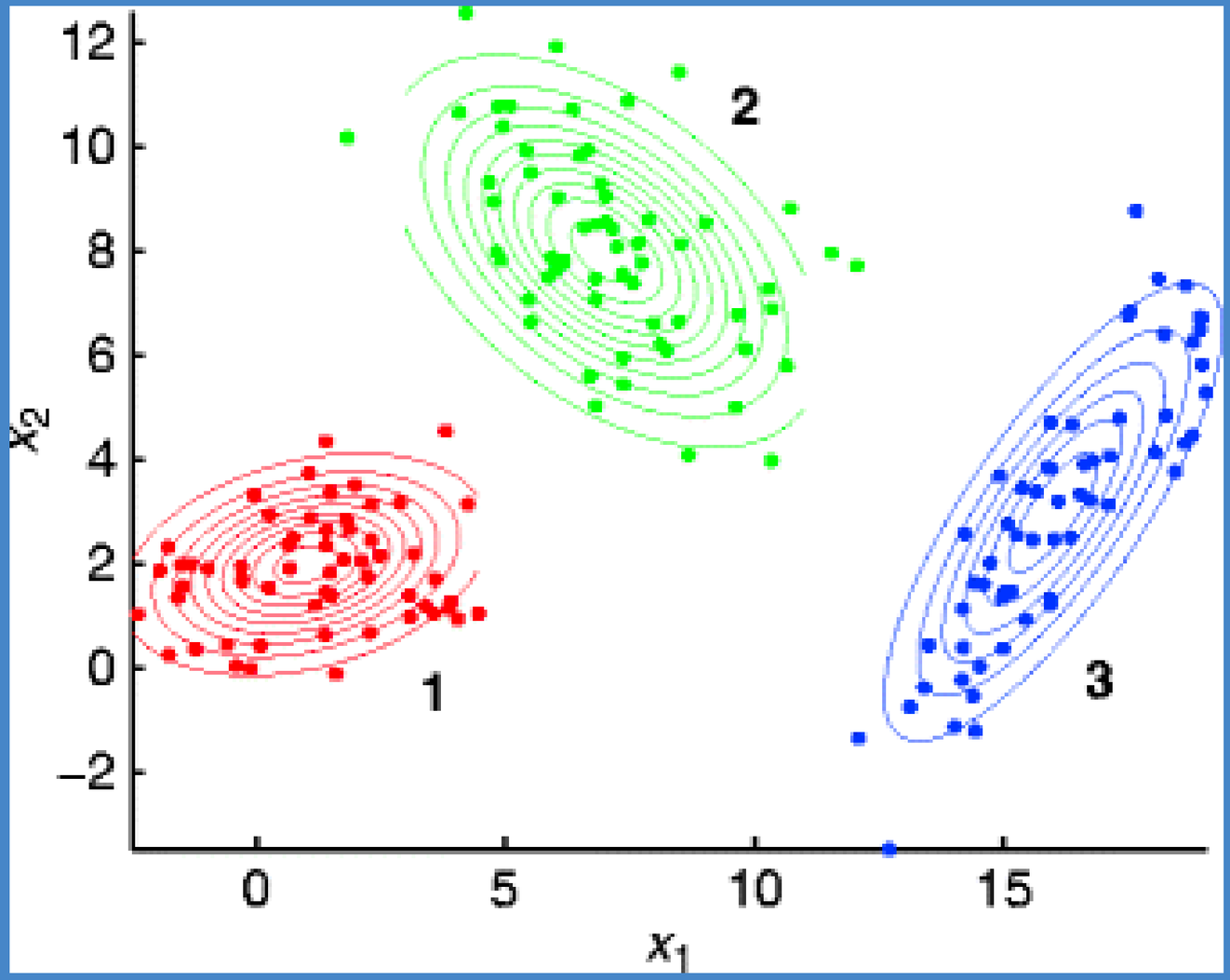
Cybersecurity, Financial Fraud, Anomaly in production sites

Matching

FAQ Bots, Job Matching, Pattern Matching

Customer Segmentation

Grouping customers based on datapoints we collect about them, creating customer persona



Underlying Concept: Similarity

Distance

If two points on a graph are close together, they are “similar”

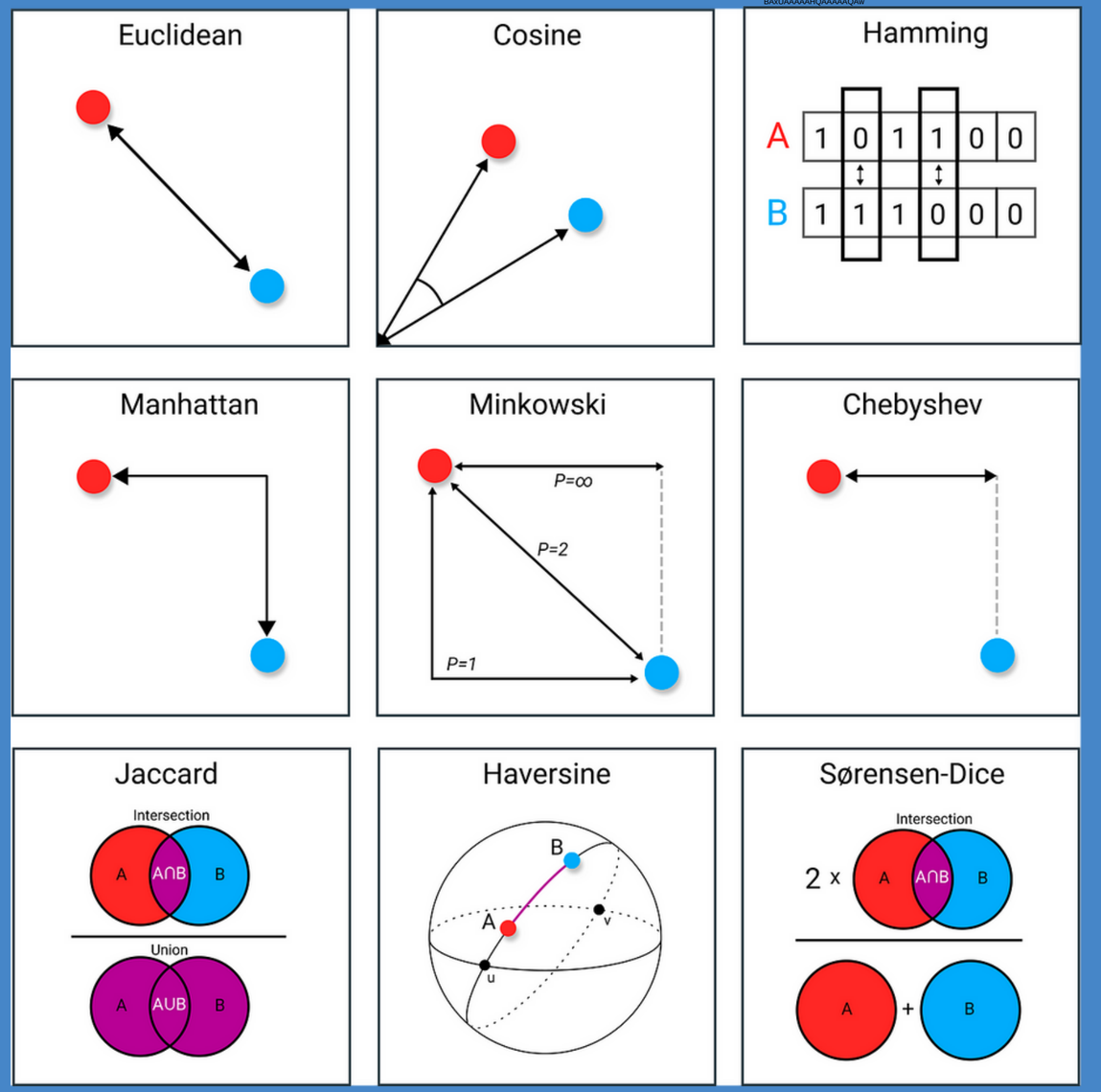
If two points on a graph are far away from each other, are they “dissimilar” or “different” from each other?

Which Similarity Measure to use?

No one perfect measure for all use-cases

Most commonly used:
For NLP tasks (vectors of numbers)- Cosine
For numerical data clustering - Euclidean
For comparing sets of recommendations - Jaccard

https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2F9-distance-measures-in-data-science-918109d069fa&psig=AOvVaw1u_JFBcrsfs7dAwhKnyMwl&ust=1697701109791000&source=images&cd=vfe&opi=89978449&ved=0CBQQ3YkBahcKEwi4xKDQi_-BAxUAAAAAHQAAAAAQAw

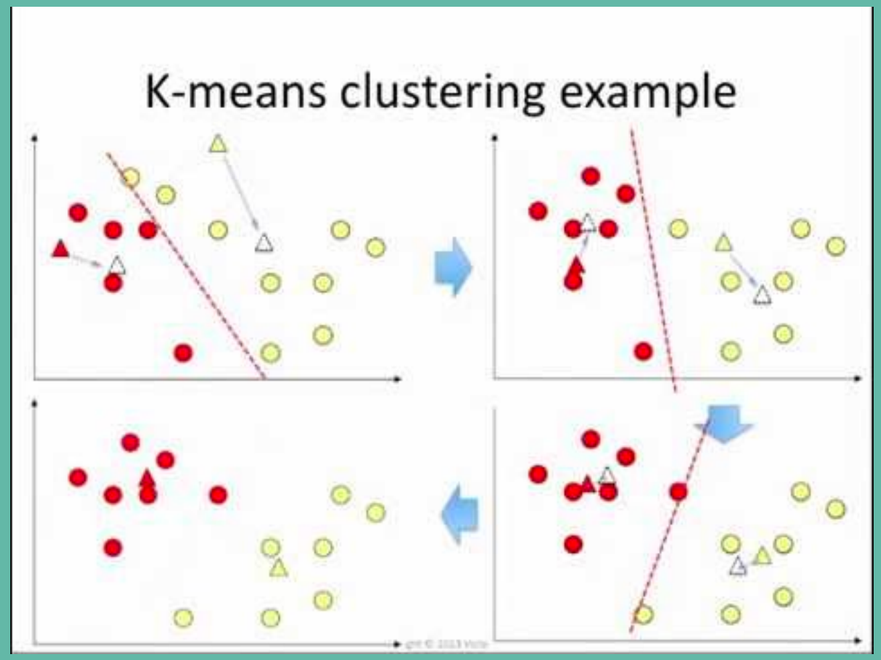


Similarity and Clustering

The closer the points are, the more similar they are

Hierarchical Clustering

Partitional Clustering



Interaction

Code Exmaple

Customer Segmentation using Unsupervised Learning (Clustering)

Link to Kaggle Notebook and Data:
https://www.kaggle.com/kavithachetanadidugu/data-analyst-mod-4-predictive-analytics

So where do we go from here?

Life-long learning

This course is the beginning of your Data Analyst path

It is not the END!

- Keep updating your knowledge base
- Read articles: Medium, KDD nuggets and many more. Just type the keyword on Google, and you get a treasure trove od resources
- Watch videos: Statquest, 3Blue1Brown, etc.
- Kaggle, Kaggle, Kaggle: Kaggle is the top community of Data Scientists and Data Analysts. There are millions of real-world datasets scraped from website, codes attempted by other people to solve the same business problems, and competitions hosted (with monetary prizes!)
- Stay curious: Always start with WHAT you are expected to do? WHY is it important? HOW can you reach the solution/final decision in the most effective way? The answers to these questions should guide your data analysis techniques, data visualisations, and your reporting

Task 2

Analyse AB test results

You are given the data collected from an AB test. An e-commerce company created a new web-page for a product, expecting customer conversion to increase. Should they implement the new web-page? Should they not? Or should they run the test longer?

Link to the data:
https://www.kaggle.com/datasets/putdejudomthai/ecommerce-ab-testing-2022-dataset1

Feedback Survey

http://www.moyyn.com/gate-feedback

Further Reading and Resources

- https://machinelearningmastery.com/random-forest-for-time-series-forecasting/ https://facebook.github.io/prophet/docs/quick_start.html https://medium.com/towards-data-science/time-series-forecasting-with-machine-learning-b3072a5b44ba https://stackoverflow.com/questions/59990884/whats-the-best-way-to-fill-the-missing-data-in-the-time-series-using-python/67571216#67571216
- Distance metrics: https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/#:~:text=Distance%20metrics%20are%20used%20in%20supervised%20and%20unsupervised%20learning%20to,Minkowski%20Distance%2C%20and%20Hamming%20Distance.
- Naive Bayes and application: https://towardsdatascience.com/text-classification-and-the-basics-of-a-naive-bayes-model-1f9096af4577
- https://medium.com/swlh/prediction-of-topics-using-multinomial-naive-bayes-classifier-2fb6f88e836f
- SVM and application: https://medium.com/jungletronics/svm-credit-card-start-to-finished-75210d644dec
- https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2

GATE

by moyyn

German Academy for Technology and Entrpreneurship

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

STARTUP SCHOOL
by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin

Der Regierende Bürgermeister von Berlin
Senatskanzlei
Wissenschaft und Forschung