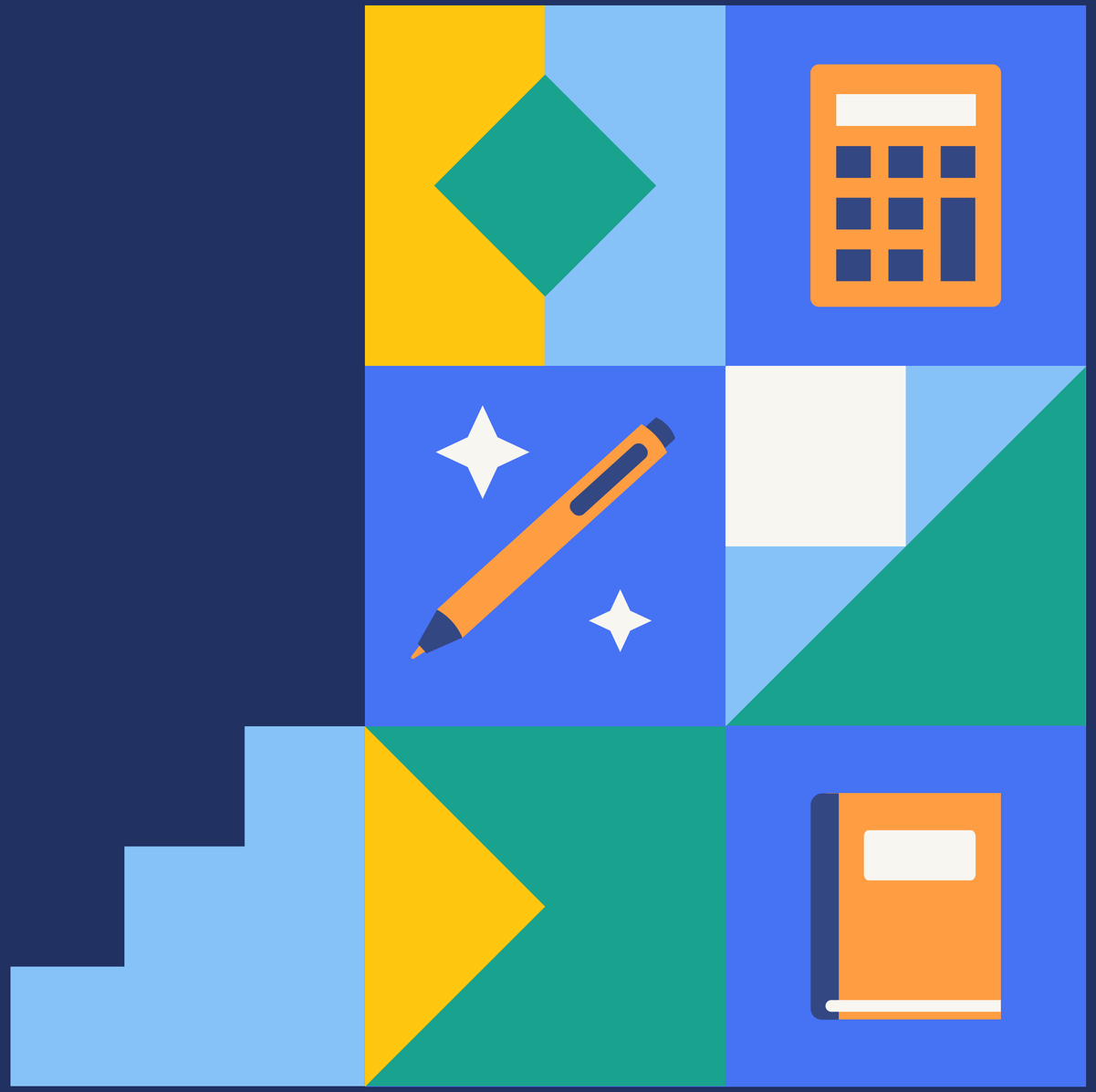
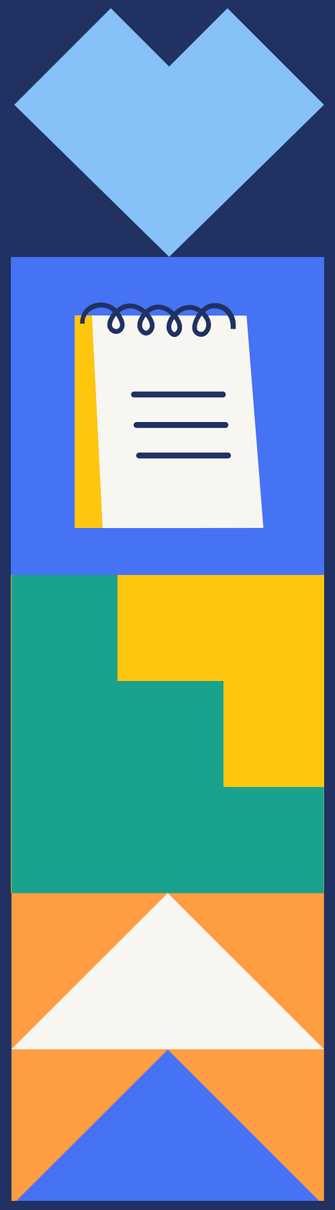
DATA 300 Recap

Faye Le, Ryan West, Maggie Byers



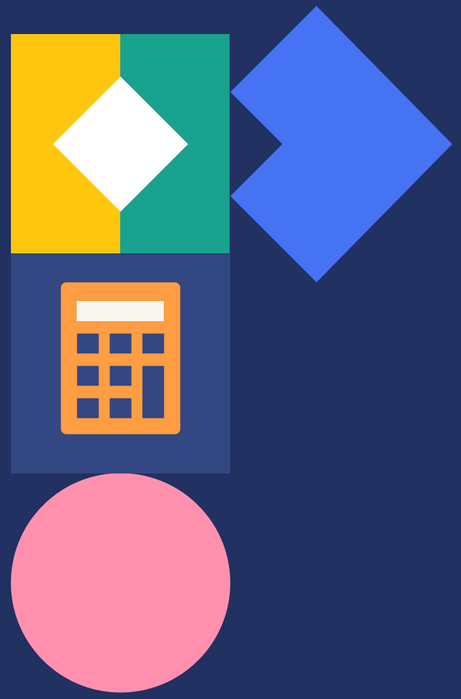


Introduction

Explored essential concepts and methods in both statistical and machine learning

Covered a variety of models crucial for understanding and analyzing datasets

- Association Rules
- Classification
- Clustering
- Regression



Association Rules

Association Rules

A machine learning technique that finds patterns and relationships between items in large datasets

Finds frequent items and items that are likely to appear in the data together

Three methods:

- Apriori
- FP Growth
- ECLAT

Association Rules

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Key Terms:
Support: number of times an item appears in a dataset
Confidence Level: percentage of times one item was bought with another

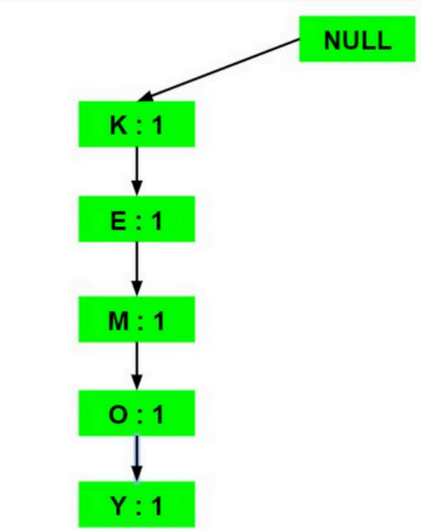
Algorithm

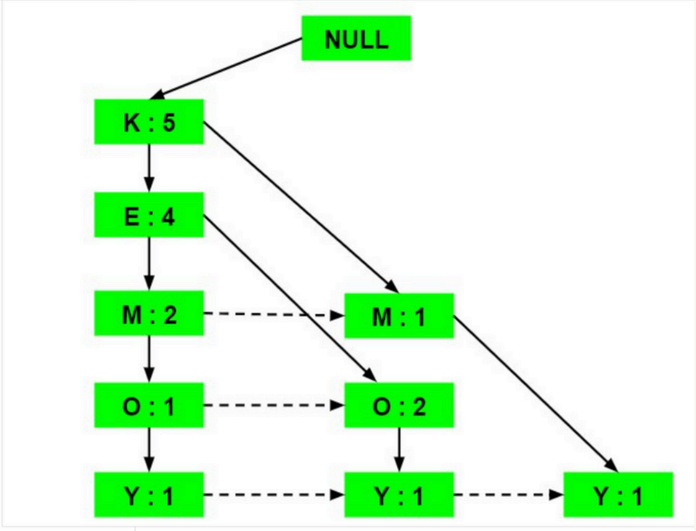
1. Set a minimum support level and a minimum confidence level
2. Starting with k=1, determine frequent items, then move to k=2...
3. Calculate confidence levels for all combinations of frequent items

Apriori Property: All subsets of a frequent itemset must be frequent;
If an itemset is infrequent, all its supersets will be infrequent.

FP Growth

Transaction ID	Items	Ordered-Item Set
T1	{E, K, M, N, O, Y}	{K, E, M, O, Y}
T2	{D, E, K, N, O, Y}	{K, E, O, Y}
T3	{A, E, K, M}	{K, E, M}
T4	{C, K, M, U, Y}	{K, M, Y}
T5	{C, E, I, K, O, O}	{K, E, O}





Items	Conditional Pattern Base	Conditional Frequent Pattern Tree
Y	{{K,E,M,O : 1}, {K,E,O : 1}, {K,M : 1}}	{K : 3}
O	{{K,E,M : 1}, {K,E : 2}}	{K,E : 3}
M	{{K,E : 2}, {K : 1}}	{K : 3}
E	{K : 4}	{K : 4}
K		

FP Growth

1. Calculate the frequency (support) of each item
2. Organize each itemset from most frequent item to least
3. Create a tree tracing each itemset
    - Start with the most frequent element
4. List the conditional pattern base and note the frequent items

ECLAT

ECLAT

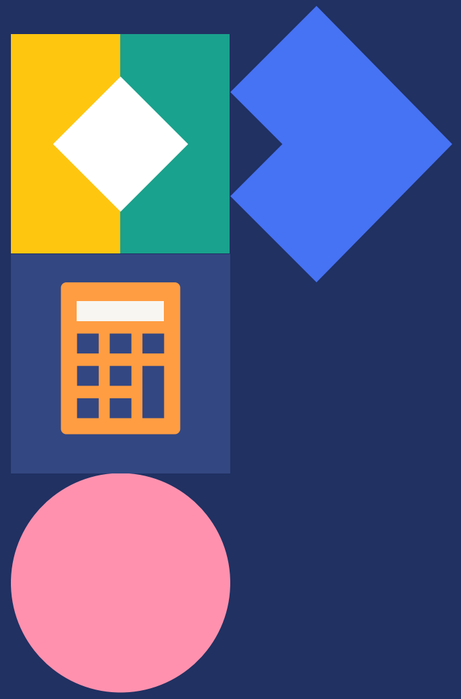
1. Item by item (k=1), list which transactions it appears in
2. Repeat for k=2, combining pairs of items in every possible combination
3. Continue doing so until no new combinations appear at that level of k

Advantages over Apriori:
- -uses less memory
- -typically faster
- -requires less computations

Item	Tidset
Bread	{T1, T4, T5, T7, T8, T9}
Butter	{T1, T2, T3, T4, T6, T8, T9}
Milk	{T3, T5, T6, T7, T8, T9}
Coke	{T2, T4}
Jam	{T1, T8}

Item	Tidset
{Bread, Butter, Milk}	{T8, T9}
{Bread, Butter, Jam}	{T1, T8}

Item	Tidset
{Bread, Butter}	{T1, T4, T8, T9}
{Bread, Milk}	{T5, T7, T8, T9}
{Bread, Coke}	{T4}
{Bread, Jam}	{T1, T8}
{Butter, Milk}	{T3, T6, T8, T9}
{Butter, Coke}	{T2, T4}
{Butter, Jam}	{T1, T8}
{Milk, Jam}	{T8}



Classification

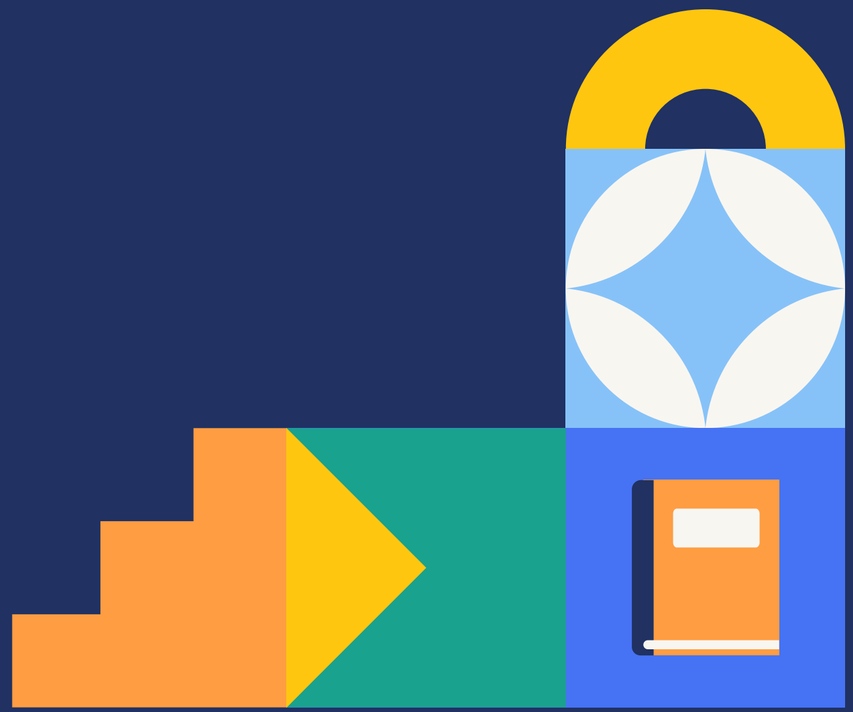
Classification

Techniques that attempt to classify each observation into a category as a way to form groups within a dataset

Goal is to be able to make accurate assumptions using the other observations in each categroy

Three methods:

- ZeroR/OneR
- K-Nearest Neighbors
- Decision Tree



ZeroR

- Simplest classifier
- Ignores all predictors
- Categorizes every observation as the majority category

OneR

- Simple but effective
- Generates one rule for each predictor in the data
- Selects the rule with the lowest error

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

Frequency Tables

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

K-Nearest Neighbors

K-Nearest Neighbors

Choose a distance metric for each predictor (0/1 for categorical, Euclidean for numeric...)

Calculate the distance between the new observation and each observation that already has the classification known

Depending on what k value is chosen, the new observation will be assigned the classficiation of the majority of its k-nearest neighbors

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

k = 3

Decision Tree

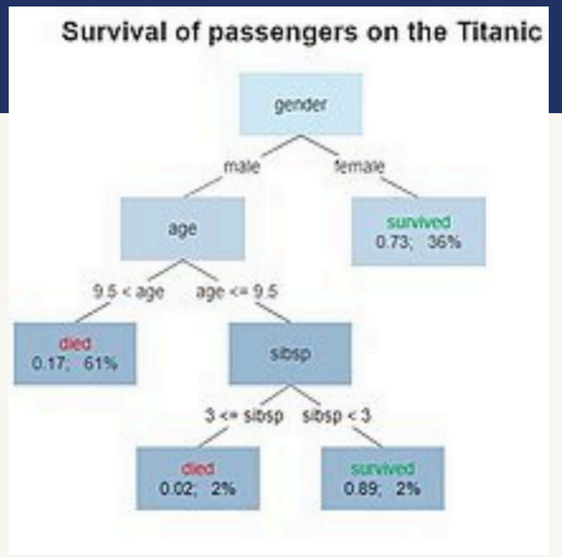
Structure

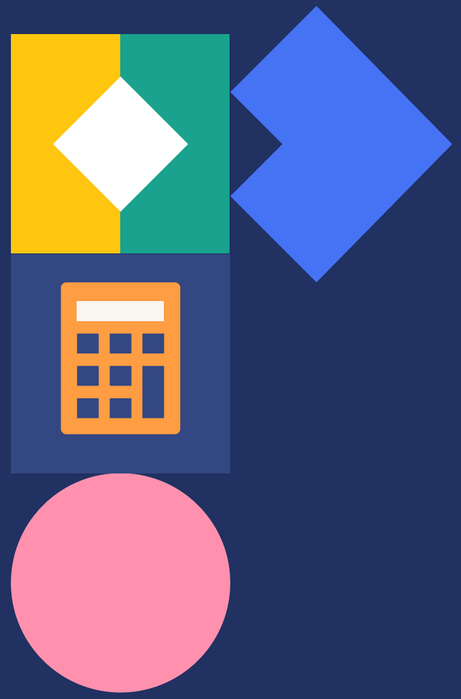
Root Node: Represents the entire dataset and the initial decision to be made.
Internal Nodes: Represent decisions or tests on attributes. Each internal node has one or more branches.
Branches: Represent the outcome of a decision or test, leading to another node.
Leaf Nodes: Represent the final decision or prediction. No further splits occur at these nodes.

Creation

Select the best predictor to split the data on based on entropy
Split the data based on classifications
Repeat until all instances in a node have the same classification

Advantages: handles nonlinear, versatility, no need for feature scaling
Disadvantages: overfitting, bias towards features with more levels, instability





Clustering

Clustering

Purpose: Groups data points into clusters based on similarity, without predefined labels.
Applications: Used in customer segmentation, image processing, and pattern recognition.
Goal: Find structures or patterns in data, where each cluster represents a group with similar features.
Clustering Techniques:

- K-Means Clustering
- K-Nearest Neighbors
- Hierarchical Clustering
- DBSCAN

K-Means Clustering

K-Means Clustering

Mechanism: Partitions data into k clusters by assigning each data point to the nearest cluster center (centroid).

Process:

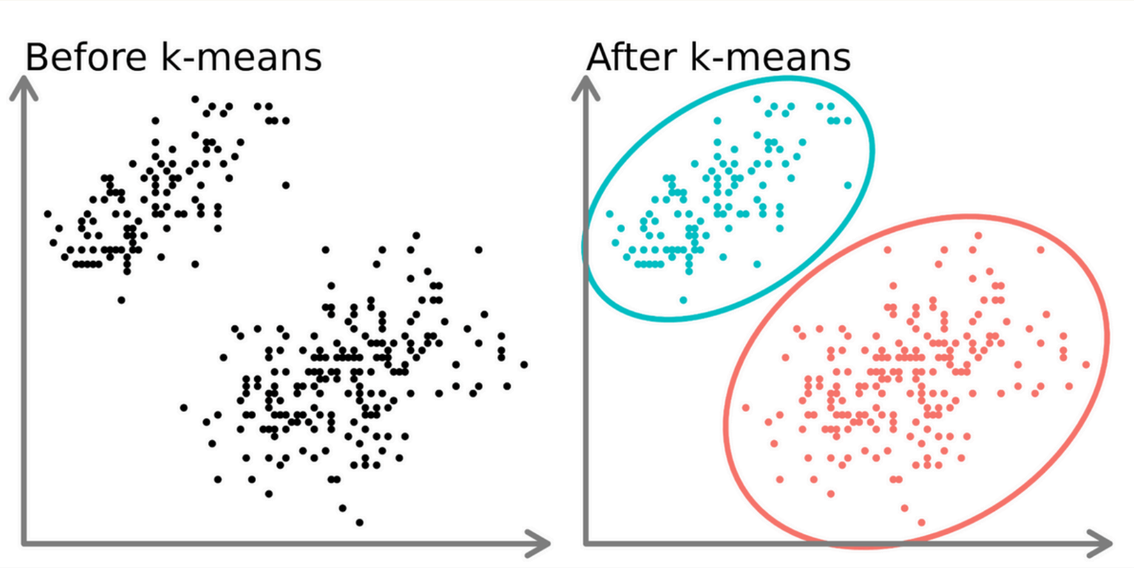
- Choose the number of neighbors, k.
- For each new point, find the k nearest points in the dataset.
- Classify the point based on the most common label among its neighbors.

Strengths:

- Efficient and works well with large datasets.
- Suitable for spherical, well-separated clusters.

Limitations:

- Requires predefining k clusters, which can be challenging.
- Sensitive to outliers and initial centroid selection.



K-Nearest Neighbors

KNN

Mechanism: A supervised learning method that can aid clustering by grouping points based on the majority class of their k closest neighbors.

Process:

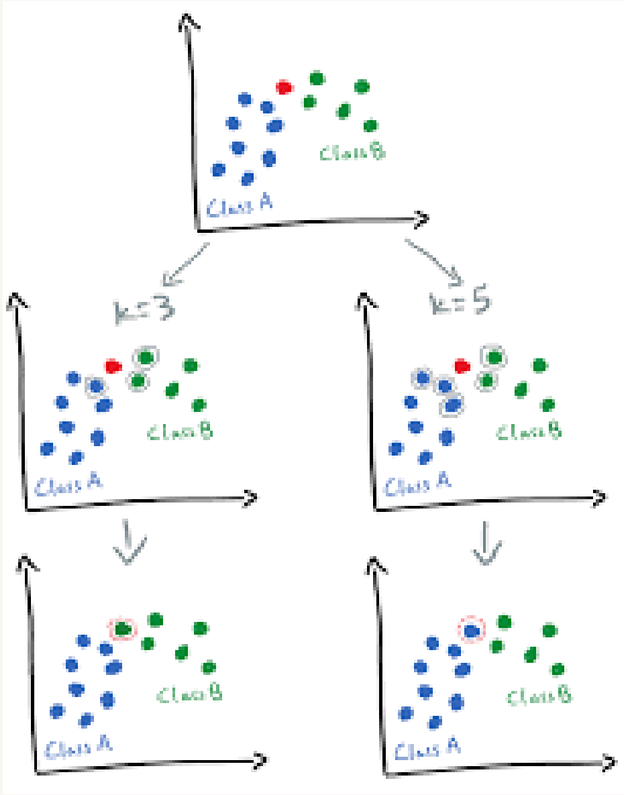
- Initialize k centroids randomly.
- Assign each point to the nearest centroid.
- Recalculate centroids by taking the mean of all points in each cluster.
- Repeat steps 2-3 until convergence (when points no longer change clusters).

Strengths:

- Simple, intuitive, and effective for clustering based on proximity.

Limitations:

- Computationally expensive with large datasets.
- Sensitive to the choice of k and the scale of data.



Hierarchical Clustering

Hierarchical Clustering

Mechanism: Builds a hierarchy of clusters that can be visualized as a dendrogram (a tree structure).

Process:

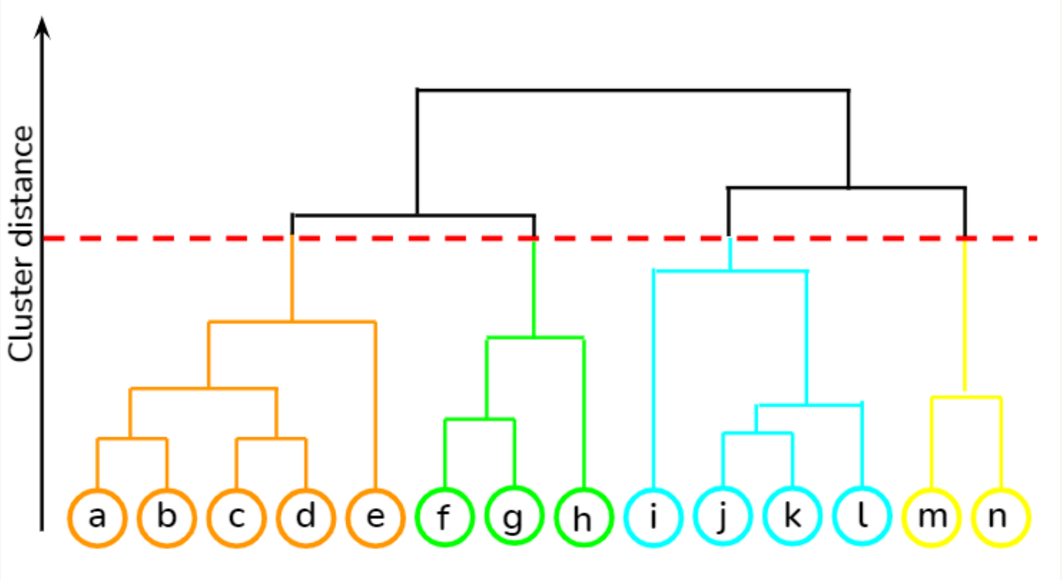
- Calculate the distance (similarity) between each pair of clusters.
- Merge the two closest clusters.
- Repeat until only one cluster remains or a stopping criterion is met.

Strengths:

- No need to specify the number of clusters upfront.
- Dendrogram allows flexibility to cut at any level for different numbers of clusters.

Limitations:

- Computationally intensive with large datasets.
- Sensitive to noise and outliers.



DBSCAN

DBSCAN

Mechanism: Groups points based on regions of high density and can identify outliers as noise.

Process:

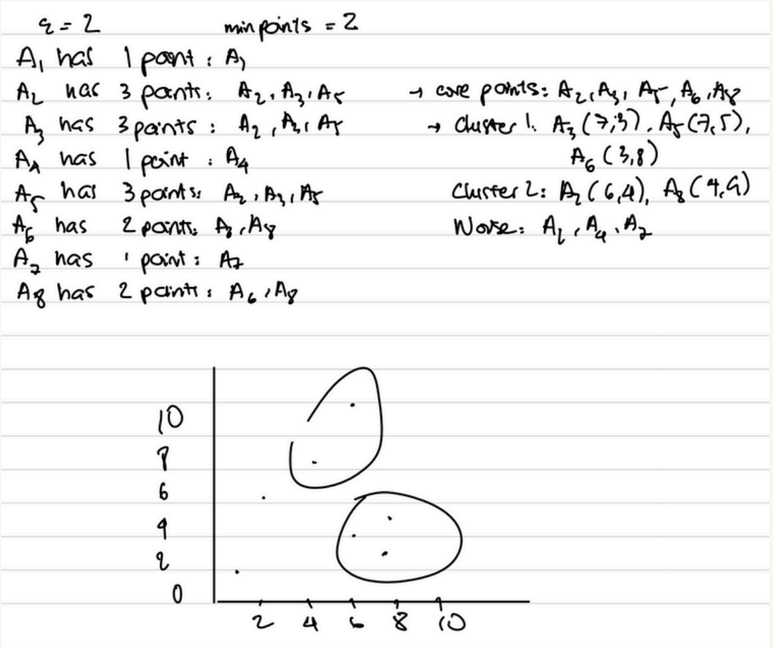
- For each point, identify neighbors within the ε radius.
- Classify points into core points (if they have at least MinPts neighbors), border points (connected to core points but with fewer than MinPts neighbors), and noise (outliers).
- Expand clusters from core points, including border points, and ignore noise.

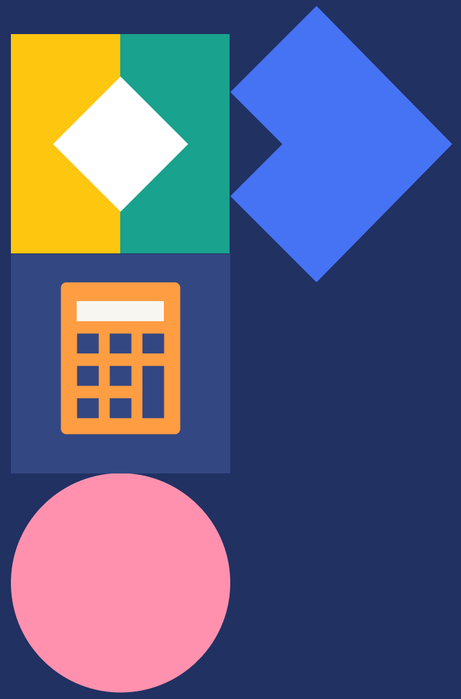
Strengths:

- Does not require specifying the number of clusters.
- Effective for clusters of arbitrary shapes and handling outliers.

Limitations:

- Results can vary depending on ε and MinPts values.
- Less effective on datasets with varying density across clusters.





Regression

Regression

A statistical method used to determine the strength of the relationships between variables, which are separated into dependent (response) and independent

Regression allows us to predict values of dependent variables based on the values of independent variables – output is continuous rather than categorical

We discussed three types:

- Decision Tree
- Linear
- K-Nearest Neighbors

Decision Tree

Decision Tree

Decision trees break down datasets based on independent variables while aiming to minimize the prediction error at every node

In DATA 300, we minimized mean squared error for each independent variable

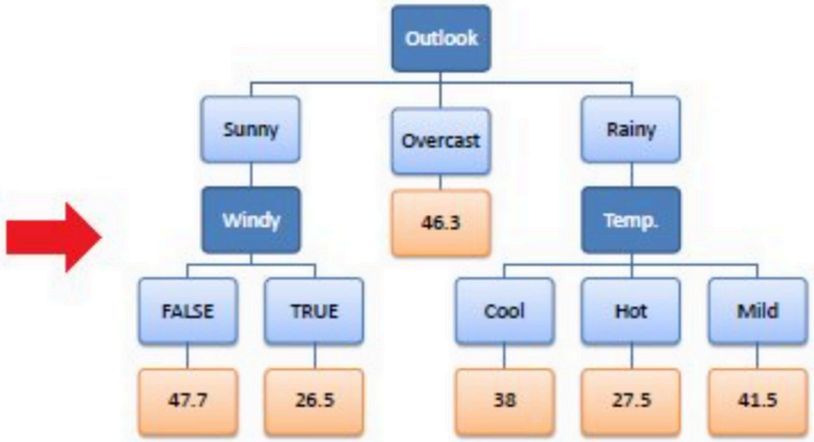
This meant that the data would be split on the variable with the smallest mean squared error

This would continue down until we reached a certain threshold, whether that be a maximum depth or a minimum number of samples

- Each leaf represents a predicted value
- Can be overfit, tree depth and minimum number of samples need to be considered

Decision Tree

Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

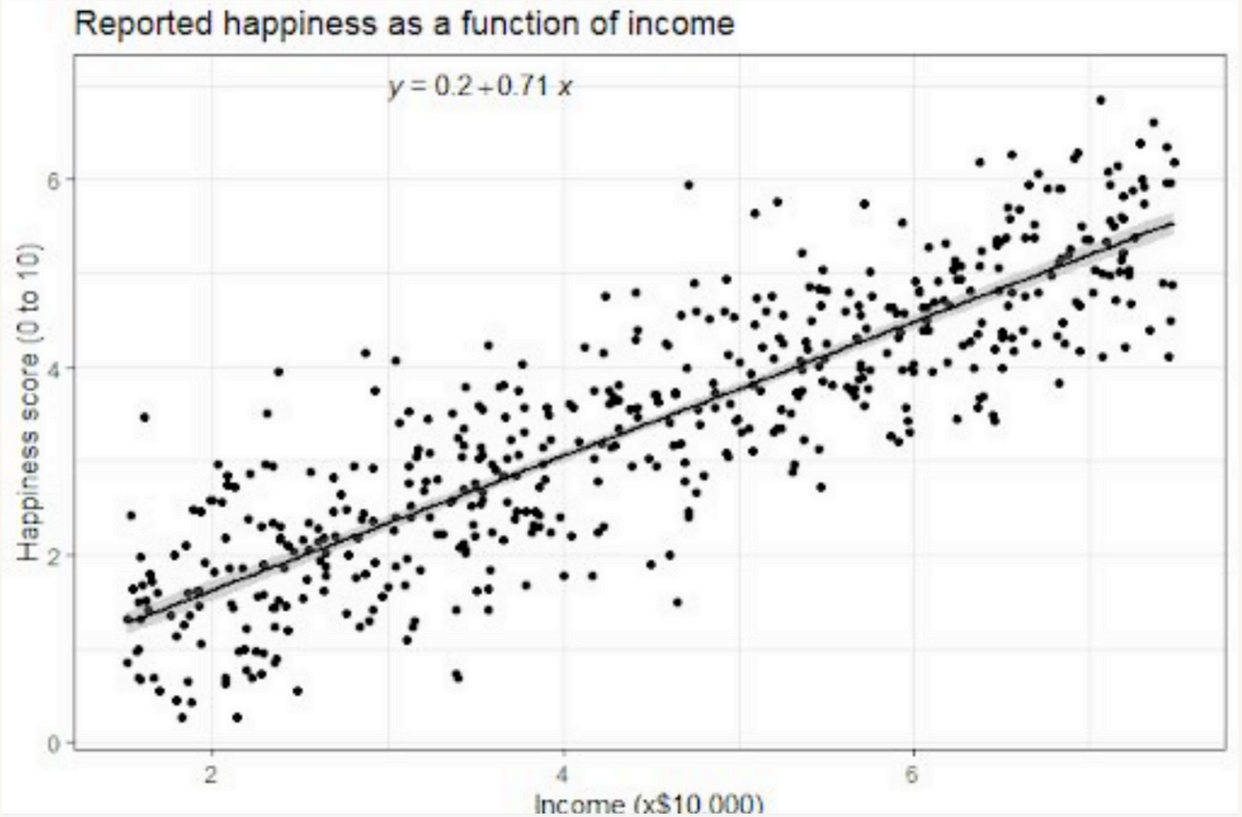


Linear Regression

Linear Regression

- Equation fits the form y = bx + a, where y is the dependent response variable, x is the independent variable, b is the slope between the points, and a is the y-intercept
- Goal is to minimize the difference between predicted vs. observed values, known as residual error
- Linear regression assumes that your dataset is linear, and all observations are independent
- Sensitive to outliers
- Only examines the relationship between two variables
- Strength of the relationship determined by the R2 value

Linear Regression



K-Nearest Neighbors

K-Nearest Neighbors

- K-NN regression uses the next closest instances in a dataset to predict a value for a new instance
- This is done by finding the distance between the attributes of each instance (In DATA 300, we used the Euclidean distance)
- The values are normalized to allow for comparison between different variables
- Choosing k: too small leads to overfitting, two large leads to over smoothing
- The final predicted value for the new instance will be the average between the k-nearest neighbors

K-Nearest Neighbors

Age	Loan	House Price Index	Distance	
25	$40,000	135	102000	
35	$60,000	256	82000	
45	$80,000	231	62000	
20	$20,000	267	122000	
35	$120,000	139	22000	2
52	$18,000	150	124000	
23	$95,000	127	47000	
40	$62,000	216	80000	
60	$100,000	139	42000	3
48	$220,000	250	78000	
33	$150,000	264	8000	1
48	**$142,000**	**?**		

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Age	Loan	House Price Index	Distance
0.125	0.11	135	0.7652
0.375	0.21	256	0.5200
0.625	0.31	231	0.3160
0	0.01	267	0.9245
0.375	0.50	139	0.3428
0.8	0.00	150	0.6220
0.075	0.38	127	0.6669
0.5	0.22	216	0.4437
1	0.41	139	0.3650
0.7	1.00	250	0.3861
0.325	0.65	264	0.3771
0.7	**0.61**	**?**	

$$X_s = \frac{X - Min}{Max - Min}$$

Discussion

Association Rules

How can you decide which rules are helpful? What could you do to prune unhelpful rules?

Classification

When might more complex algorithms give a result like ZeroR?

Clustering

Why might it be possible for clusters to produce misleading results? How could you prevent this from happening?

Regression

Sometimes, regression models may unintentionally reinforce biases. How could you mitigate potential bias in your model?

Thank you for your listening