Machine Learning Markdown

Apriori Algorithm

Definition:

- Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules.
- The parameters “support” and “confidence” are utilized.
- Support = items’ frequency of occurrence Confidence = conditional probability

How it works:

- Items in a transaction = item set
- Algorithm identify frequent, individual items (items with higher frequency than the support)
- Expands analysis to larger frequent itemsets

Example:

- Support* = 3, confidence = 80%
-

Transaction ID	Items
T1	I1, I2, I3, I4
T2	I2, I3
T3	I3, I4
T4	I2, I3, I4

- I2 with I3, confidence = 3/3 = 100%

Data Preprocessing:

- Connect to RDS
- Load pivot table (product_ids as headers)
- Set item association function

Feature Selection:

- Understand items brought by the same customer can increase conversion rates in ecomm and drive revenues growth (cross sell)

- Apriori algorithm is popular for this type of analysis
- Apriori gives confidence level of recommendation that helps data analysts decide the right threshold for website recommended products

Model Selection:

- Benefits
  - Most simple algorithm among association rule learning
  - Broadly adopted for basket analysis
  - Easy to understand and interpret
  - Exhaustive: finds all rules with confidence levels
- Limitations
  - Not good for small datasets
  - Takes time to run

K-Means Cluster Analysis

Data preprocessing:

- Connect to RDS
- Replace NaN values with zeros
- Drop columns (product categories) not needed
- Scale the data
- Customer_id as index

Feature Selection:

- Create customer segmentation based on product category
- Goal is to target specific segments based on categories purchased
- Unsupervised model, since there's no dependent variable (Y)
- No need to split and train the data for unsupervised model

Model Selection:

- Benefits
  - Simple to implement
  - Runs relatively quickly
  - Can scale large datasets
- Limitations
  - Sensitive to outliers
  - User defines # of clusters

- Hard to interpret output since there's no Y variable

Latent Dirichlet Allocation (LDA) Machine Learning Model

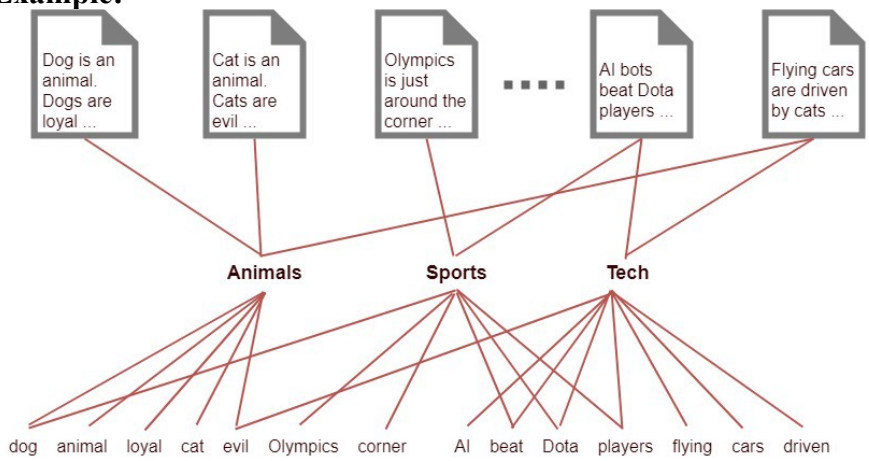
Definition:

- Topic modelling is the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent)
- LDA is a popular model in topic discovery.

How it works:

- Fixed number of topics
- Each topics represents a group of words
- LDA maps all documents to this topic

Example:



Data Preprocessing:

- Connect to RDS
- Remove unwanted characters, numbers, symbols and stop words
- Remove non value table
- Separate data between 1 star and 5 star reviews
- Use nlp to remove worlds are not aggregating (only noun and adjectives)

Feature Selection:

- Topic discovery for customer reviews to gather feedback and identify themes: product qualities and what has to be improved
- Find 'relevant' topics and identify trends

  Topic Modelling is an unsupervised approach used for finding and observing the bunch of word

Model Selection:

- Benefits
  - Largely used for topic discovery
  - Simple to implement
  - Runs relatively quickly
  - Probabilistic model
- Limitations
  - User defines # of topics
  - Hard to interpret output since there's no Y variable