the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

Unsupervised Learning

Luiza Sayfullina, SILO AI

Supervised Learning

Student: makes the predictions

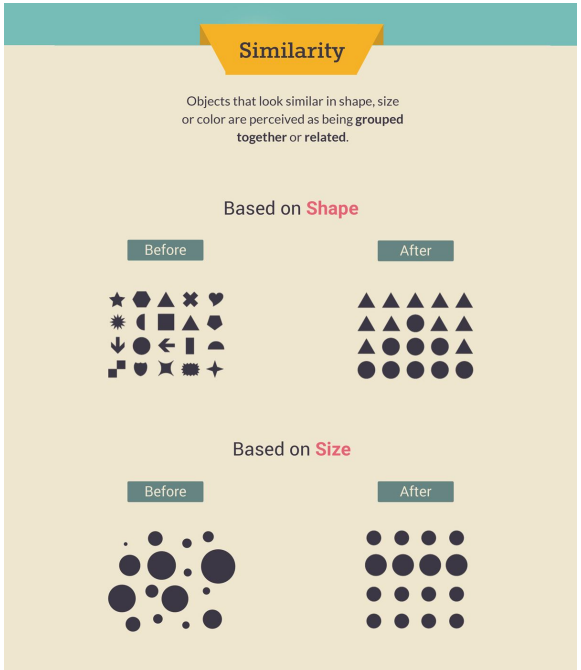
Teacher: knows the real answers, corrects the student in case he or she is wrong

Unsupervised Learning

Student: makes the predictions based on the data

Teacher: knows the real answers, corrects the student in case he or she is wrong

Grouping objects in an unsupervised way



Data Engineering

Concern of Unsupervised learning

$$P(X, Y) = P(Y|X)P(X)$$

Concern of Supervised learning

Bayes formula

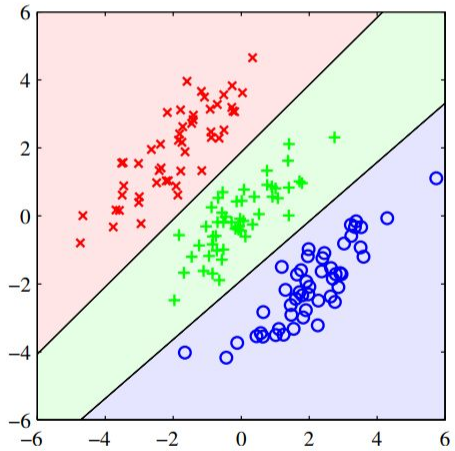
Student: learns without teacher just by having an access to all data points

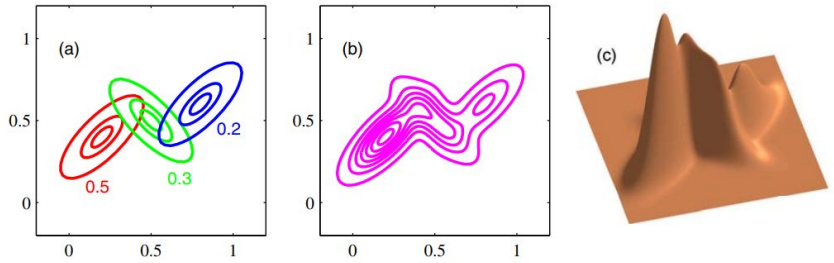
Teacher: knows the real answers, corrects the student in case he or she is wrong

Unsupervised Learning

OBJ The goal of Supervised Learning is to learn to separate data P(Y|X)
The measure of success due to the presence of targets is easy to pick.

OBJ The goal of Unsupervised Learning is to learn from data density P(X), often hard to measure the performance of a unsupervised approach



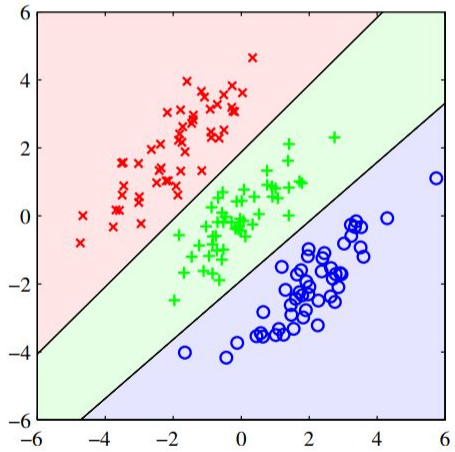


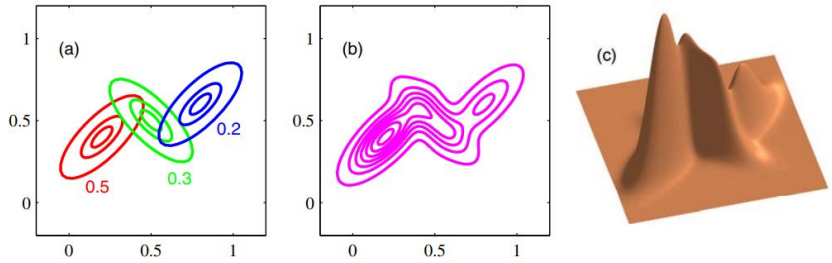
Question: How 3D data is approximated in above figure?

Unsupervised Learning

The goal of Supervised Learning is to learn to separate data P(Y|X)
The measure of success due to the presence of targets is easy to pick.

The goal of Unsupervised Learning is to learn from data density P(X), often hard to measure the performance of a unsupervised approach

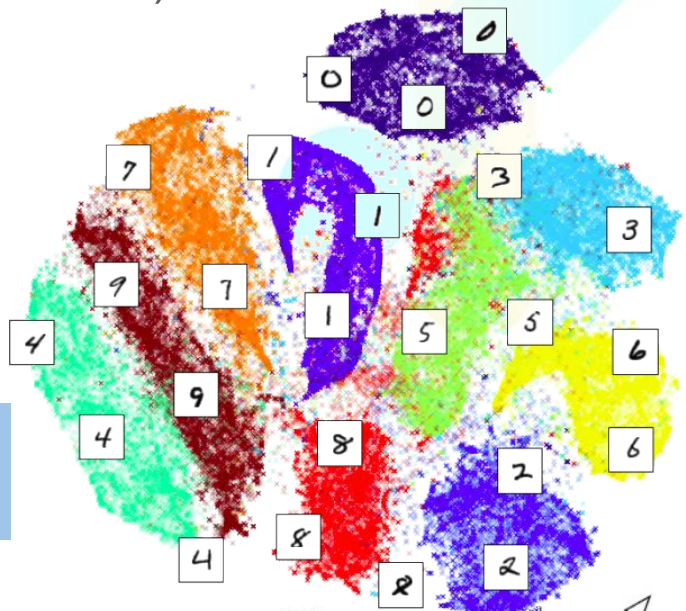




The data in 3D space was approximated by a mixture of 3 gaussian functions.

Unsupervised Learning

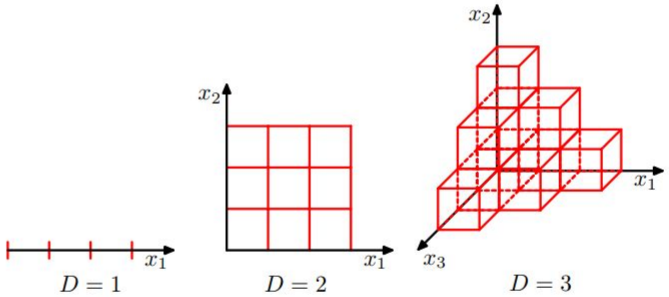
- Learning without a teacher from the data
- Try to learn from data distribution P(X) (its structure) and infer some conclusions about the data



Hand-written digits shown in 2 space have a very good separation, showing some structure.

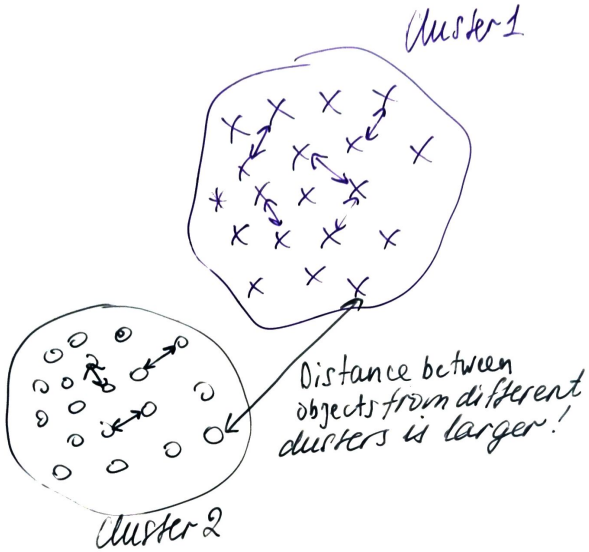
Unsupervised Learning

- Learning without a teacher from the data
- Try to learn from data distribution P(X) about its structure and infer some conclusions about the data
- If the dimension of data is less than 4, then one can estimate density of data
- However with more variables estimating density is hard (also known as curse of dimensionality)
- Therefore for high dimensions we aim at describing X values (data) with a set of regions where density is large.



Cluster analysis

- The goal is to find multiple convex regions of X space that contain P(X) modes.
- Group objects into subsets or clusters such that objects within each cluster are more closely related to one another than objects assigned to different clusters.
- An object can be described by a set of measurements or by relation to other objects.
- Clustering analysis is based on the similarity or dissimilarity degree between the objects being clustered



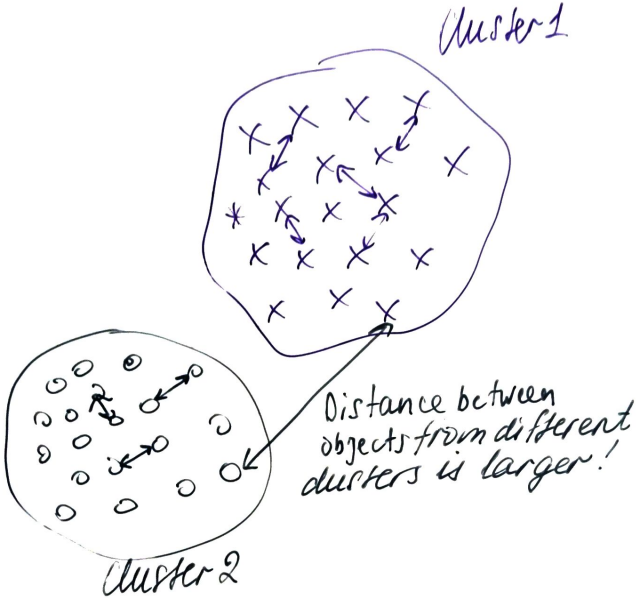
Types of clustering algorithms

- Combinatorial algorithms
- Mixture modelling
- Mode seeking

Proximity measures

- The input to clustering algorithm should be dissimilarity matrix which defines dissimilarity between each pair of objects
- Dissimilarity measure has a greater importance than a clustering algorithm itself
- Dissimilarity between objects *i* and *k* is measured as the sum of distances between coordinates :

$$D(x_i, x_k) = \sum_{j=1}^{P} d_j(x_{ij}, x_{kj})$$



Distance measures for numerical data

- Squared distance:

$$d_j(x_{ij}, x_{kj}) = (x_{ij} - x_{kj})^2$$

- Absolute distance:

$$d_j(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}|$$

- Correlation measure

$$d(x_i, x_k) = corr(x_i, x_k)$$

Distance measures for ordered categorical features

- Categorical features describe the object with a category name
- Ordered categorical features can be naturally assigned integer values that will reflect the order
- E.g. the grades A, B, C, D can be mapped to integer numbers like 5, 4, 3, 2
- In comparison categorical features in general are not trivial to compare but possible.

Question: How would you measure the distance between two city names?
What about colors?

K-means clustering

- Clustering aims at assigning N observations to K clusters
- K-means clustering aims at minimizing *within cluster variance*, meaning that the sum of distances between points within their clusters in minimized:

$$\sum_{k=1}^{K} \sum_{c(i)=k} \sum_{c(j)=k} ||x_i - x_j||^2$$

- The name comes from the fact that we want to receive K clusters which would have the centers as the mean of all points in that cluster
- The algorithm is based on the Euclidean distance

K-means algorithm

1. Randomly assign the centers of K clusters (*cluster centroids* = cluster centres)
2. Assign each sample to the closest cluster centroid according to the Euclidean distance
3. Recompute the centroid as the mean of all points for each cluster
4. Repeat steps 2 and 3 until points are re-assigned to the new cluster

Question: does algorithm provide locally or globally optimal solution?

K-means animation

K-means algorithm

1. Randomly assign the centers of K clusters (*cluster centroids* = cluster centres)
2. Assign each sample to the closest cluster centroid according to the Euclidean distance
3. Recompute the centroid as the mean of all points for each cluster
4. Repeat steps 2 and 3 until points are re-assigned to the new cluster

The method provides suboptimal local minimum and due to randomness in step 1 the algorithm can give different cluster assignments between runs

K-means animation

K-means: practicalities

- The centers of clusters can be initialized with random samples from the data
- The number of clusters can
  - Come from the task (e.g. we want to divide people into similar backgrounds to the groups of the same size, meaning the number of groups is predefined)
  - Should be based on the data and automatically calculated, like Elbow (using inertia measure in sklearn) and Silhouette methods
  - Silhouette method requires running K-means for various K and picking the K with the highest score (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)

Silhouette Score

Compares how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Helps to choose the number of clusters.

For data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

For each data point $i \in C_i$, we now define

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Task: prove that Silhouette score lies in the interval of [-1, 1]

The better the clustering, the higher the mean Silhouette score

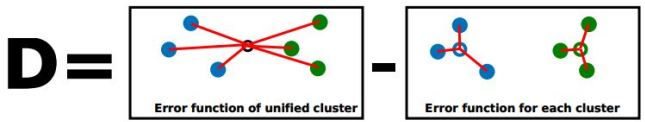
Silhouette Score: practicalities

- Run K-means for different number of clusters
- For each number of clusters, evaluate mean Silhouette score and pick the one where the Silhouette score is the highest

Methods that do not require setting a cluster number … but require setting other parameters

- Agglomerative Clustering (Hierarchical method)
  - Recursively merges the pair of clusters that minimally increases a given *linkage* distance
  - **Linkage** defines the distance between clusters
    - **Single linkage**: the distance between two clusters is the shortest distance between two points in each cluster
    - **Complete linkage:** the distance between two clusters is the longest distance between two points in each cluster
    - **Average linkage:** the distance between clusters is the average distance between each point in one cluster to every point in other cluster
    - **Ward linkage:** sum of distances of each point to global centroid minus sum of distances of each point to cluster centroids for all clusters.

Ward linkage:



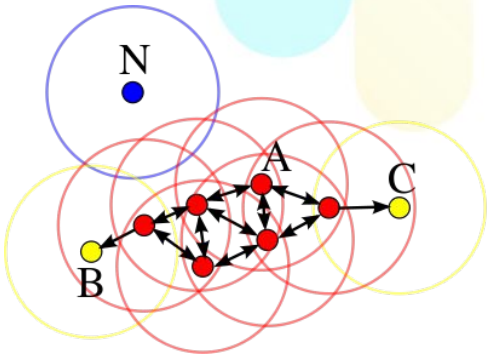
The distance (D) between to clusters is defined as the error function of the unified cluster minus the error functions of the individual clusters

Methods that do not require setting a cluster number … but require setting other parameters

- DBScan (Density-based method)
  - Full name: *Density-Based Spatial Clustering of Applications with Noise.*
    Finds core samples of high density and expands clusters from them.
  - Requires nearest neighbour computation, one might prefer fast OPTICS implementation over sklearn
  - Pros: Can sort data into clusters of varying shapes
  - Pros: Is great with handling outliers in the dataset
  - Cons: struggles when the density of data is different
  - Cons: suffers from high dimensionality

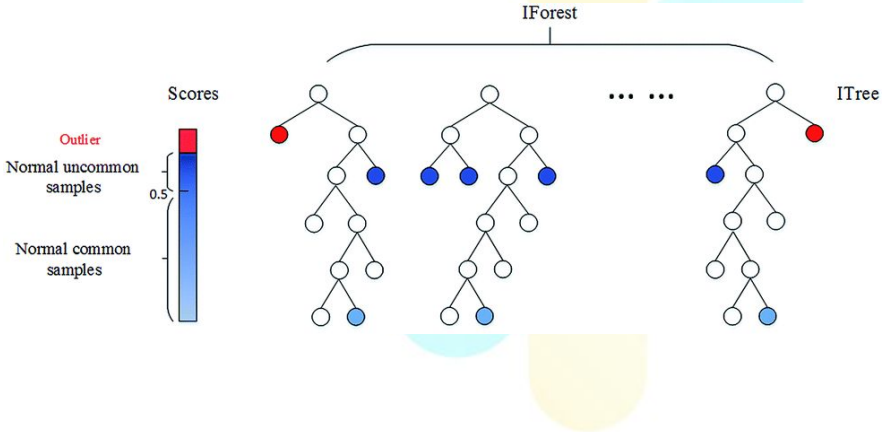
Core points have at least MinPs points within eps distance

A-core point, B,C - border points and N-outlier point



Methods that do not require setting a cluster number … but require setting other parameters

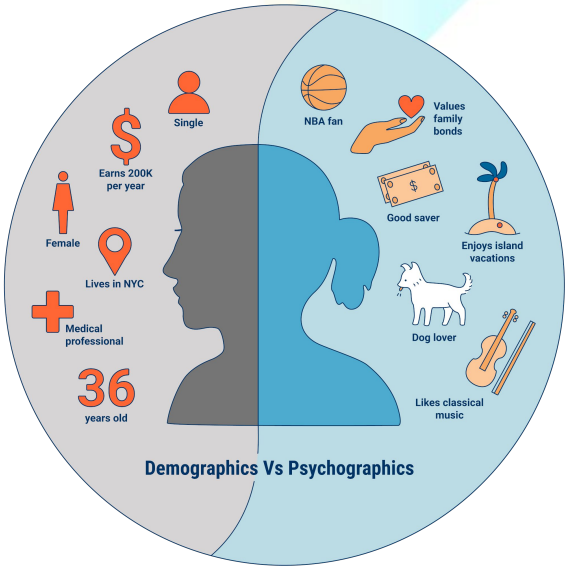
- Isolation Forest
  - Builds a binary tree where at each node the features splits randomly
  - Outlier points tend to have less splits before reaching the leaf, meaning they differ more than other points.
  - Isolation Forest outputs abnormality score, the more it is, the more abnormal is the point



Unsupervised learning use cases

- Grouping entities based on their similarity:
  - Segment your clients according to their preferences in buying and spending their money and design a strategy for each group

Discuss with peers how you could segment
your clients (which features you would use) if you would work in AI consulting company and your client will be an organization?



Unsupervised learning use cases (NLP)

- Create embeddings
  - From large corpus of text create word embeddings (vectors) using a principle *“words are described by their neighbours”*
    “Tell me who your friends are, and I will tell you who you are”

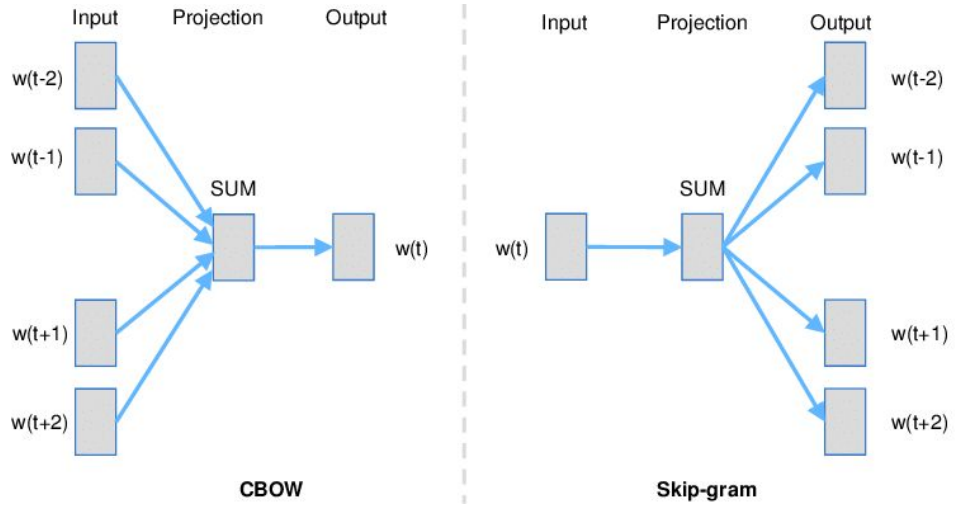
- *The company X has made a large investment deal last week.*
- *X hired 5 Data Scientists during half of the year.*
- *X released its first product which recognizes human speech with a high accuracy.*
- *The CEO of X has announced opening new offices in London.*

What can you infer about the word X?

Unsupervised learning use cases (NLP)

- Create word embeddings (vectors) using Neural Networks
  - From large corpus of text create word embeddings using a principle *“words are described by their neighbours”.*

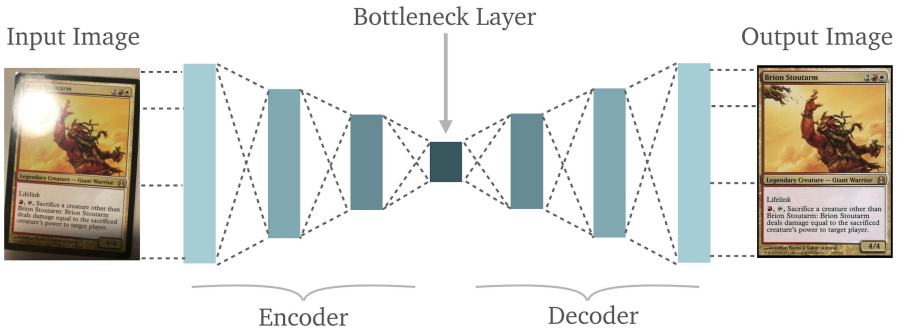
Continuous bag-of-words model predicts the missing word in the middle



Skip-gram model predicts neighbouring words from the left and right from the word

Unsupervised learning use cases (NLP)

- Unsupervised pre-training of Neural Networks
  - Train a network (e.g. auto-encoder) to denoise an image or remove some unwanted effects and uses encoder weights as initial weights for another network that solves a supervised task
  - Training data for denoising task is based on samples only.
  - Why denoising network learns Something useful about the data?



Why Does Unsupervised Pre-training Help Deep Learning? Erhan et al., 2010
http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

Unsupervised learning use cases (NLP)

- Language modelling
  - Aims to compute the probability of a sequence of words or conditional probability of a word to follow a given sentence
  - Using Chain Rule the probability of a sequence of words can be calculated as (traditional approach):
    $P(W) = P(w1, w2, w3, w4, \ldots, wn) = P(w1)\, P(w2|\, w1)\, P(w3|w1, w2) \ldots P(xn|\, x1, \ldots, xn-1)$
  - These days Neural Network-based models are trained to predict the next word or predict masked words in the sentence (BERT model).
  - Language modelling again does not require anything rather than a big corpus of language data.

P (person | Are you a cat or a dog) > **P** (ago | Are you a cat or a dog)
P (Are you a cat or a dog person) > **P** (Dog person you are a cat or a)

Unsupervised learning use cases (Anomaly detection)

- Underlying method: clustering, e.g. DBScan or Agglomerative Clustering
- Assumption: points lying outside of main clusters or forming very small clusters are the candidates for anomalies

Unsupervised learning use cases (Contract clause templates)

- Underlying method: Agglomerative Clustering on word2vec-based clause representation
- Assumption: Contract clause template is the most representative sample from each clause clusters with a specific information removed

Unsupervised learning use cases (Frequently bought products)

- Underlying approach: Association rules using Apriori algorithm (IF-THEN rules)

Frequently bought together

+

Total price: $35.78

Add both to Cart

Add both to List

Association rule 2: Support 13.4%, confidence 80.8%, and lift 2.13.

$$\begin{bmatrix} \text{language in home} & = & \textit{English} \\ \text{householder status} & = & \textit{own} \\ \text{occupation} & = & \{\textit{professional/managerial}\} \end{bmatrix} \Downarrow \text{income} \geq \textit{\$40,000}$$

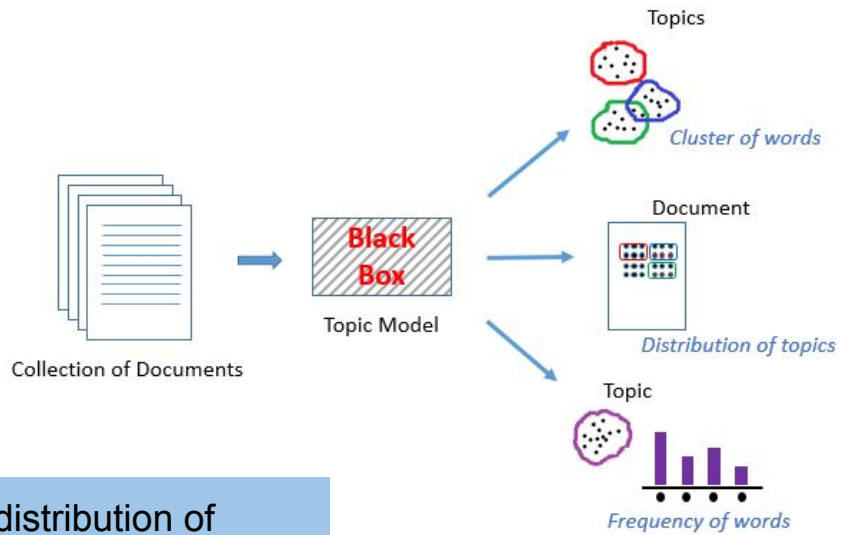
Association rule 3: Support 26.5%, confidence 82.8% and lift 2.15.

$$\begin{bmatrix} \text{language in home} & = & \textit{English} \\ \text{income} & < & \textit{\$40,000} \\ \text{marital status} & = & \textit{not married} \\ \text{number of children} & = & 0 \end{bmatrix} \Downarrow \text{education} \notin \{\textit{college graduate, graduate study}\}$$

Unsupervised learning use cases (Topic modelling for emails)

- Underlying approach: Latent Dirichlet Allocation (a probabilistic model)
- Assumption: Each email / document consists of set of topics. Each topic can be modelled with a set of words.

As a result document represented as vector where each dimension represents a topic presence. Documents can be clustered based on this vector.



Why topic explained by a distribution of unordered words can be partially unreliable?

Conclusions

- Unsupervised learning aims to learn from the data about its distribution
- Tasks falling under unsupervised learning are not limited to:
  - Clustering
  - Association Rules
  - Density Estimation (finding the areas where the data points are located)
  - Anomaly Detection
  - Word Embeddings
- K-means is a classical clustering algorithm aiming to decrease within cluster variance, but for real tasks consider using as well other approaches such as DBScan and Agglomerative Clustering.