Review of Data Mining Algorithms

Supervised learning

- Classification/prediction

Unsupervised learning

- Clustering
- Association rule mining

Semi-supervised learning

- Active learning

Recommender systems

- Collaborative filtering
- Matrix completion

Graphical models

Terminologies

- Dataset
  - Training set
  - Testing set
  - Validating set
- Data representation
  - Feature vector
- TF/IDF

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{idf}(\text{this}, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\text{tf}(\text{example}, d_2) = 3$$

$$\text{idf}(\text{example}, D) = \log \frac{2}{1} \approx 0.3010$$

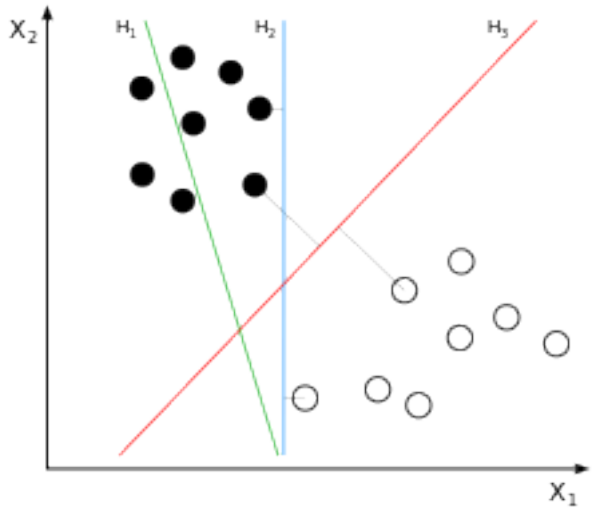
$$\text{tfidf}(\text{example}, d_2) = \text{tf}(\text{example}, d_2) \times \text{idf}(\text{example}, D) = 3 \times 0.3010 \approx 0.9030$$

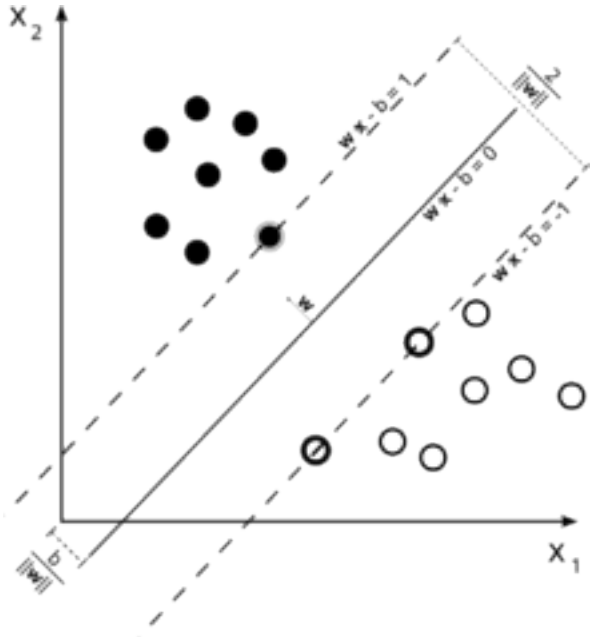
Supervised Learning

- Regression
  - Linear regression
  - Logistic regression
- Naïve Bayes
  - Strong independence assumption
- K-nearest neighboring (KNN)
- Decision Tree
  - C4.5
  - Can handle both numerical and categorical features
  - Missing values

Support Vector Machine

- Find a hyper-plane to maximize the functional margin.





- Evaluation
  - ❑Accuracy
  - ❑Precision-recall

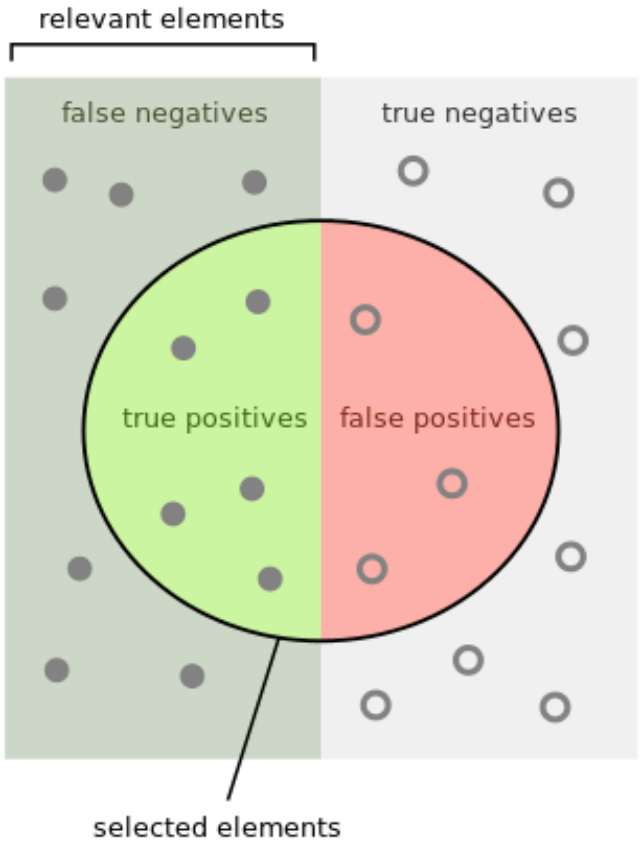
$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

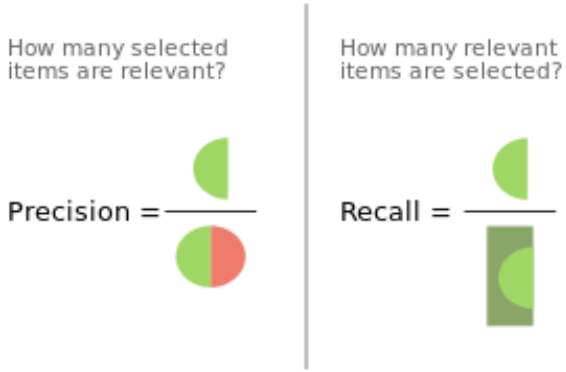
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

  - ❑F1 score

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

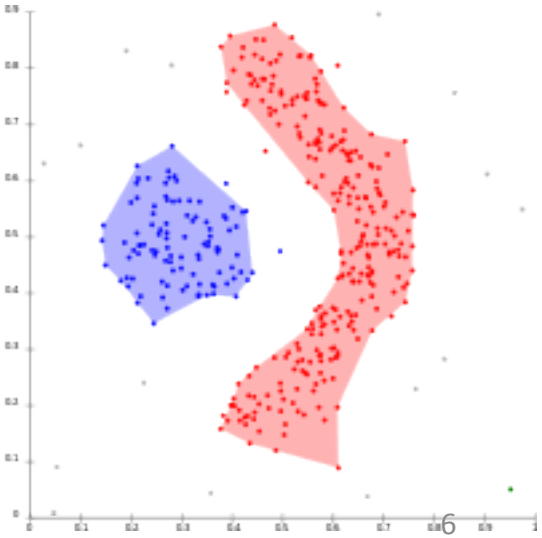
- Over fitting
  - ❑Cross-validation
  - ❑Regularization: L1/L2-norm
  - ❑Early stopping
  - ❑Pruning



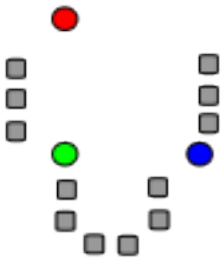


Unsupervised Learning

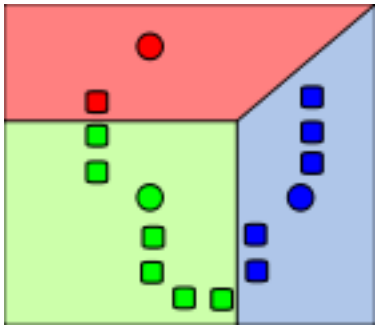
- Clustering
  - ❑K-means
  - ❑Spectral clustering
  - ❑Hierarchical clustering
  - ❑Density-based clustering (DBSCAN)
- Distance metric
  - ❑Euclidean
  - ❑Manhanttan
  - ❑Cosine
  - ❑…



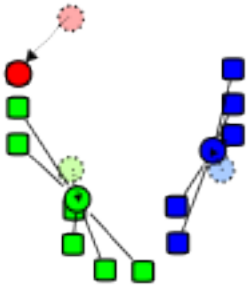
K-Means



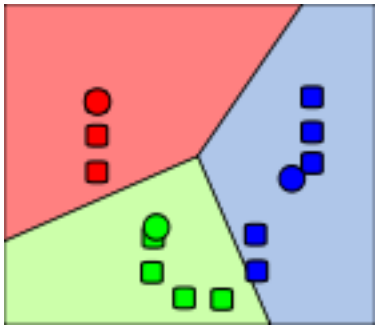
1) k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).



2) k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

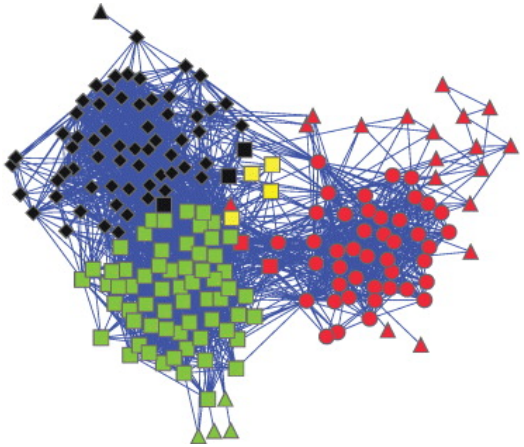


3) The centroid of each of the k clusters becomes the new mean.



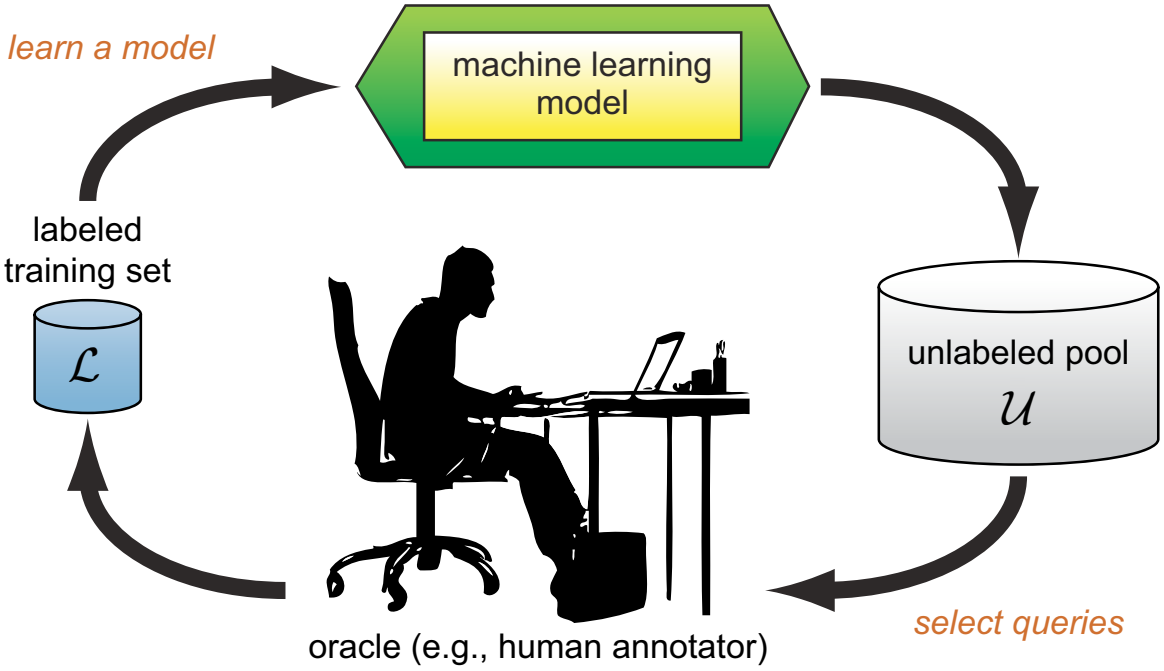
4) Steps 2 and 3 are repeated until convergence has been reached.

- Association rule mining (market basket analysis)
  - {x,y} =>{z}
  - {x,y,z}=>{u,v}
  - …
- Graph-based community detection
  - Modularity maximization-based
    - Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

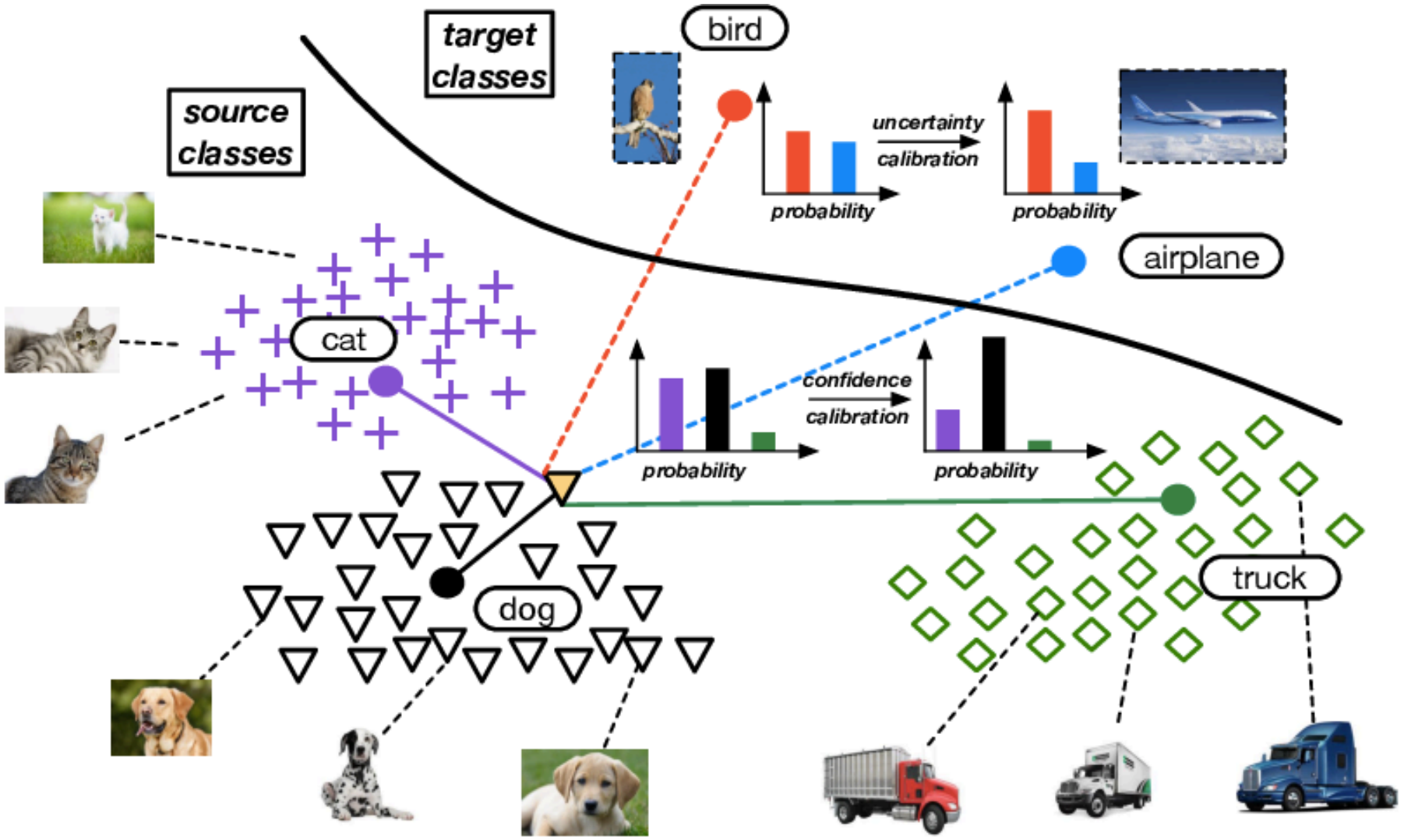


Semi-supervised learning

- Active learning

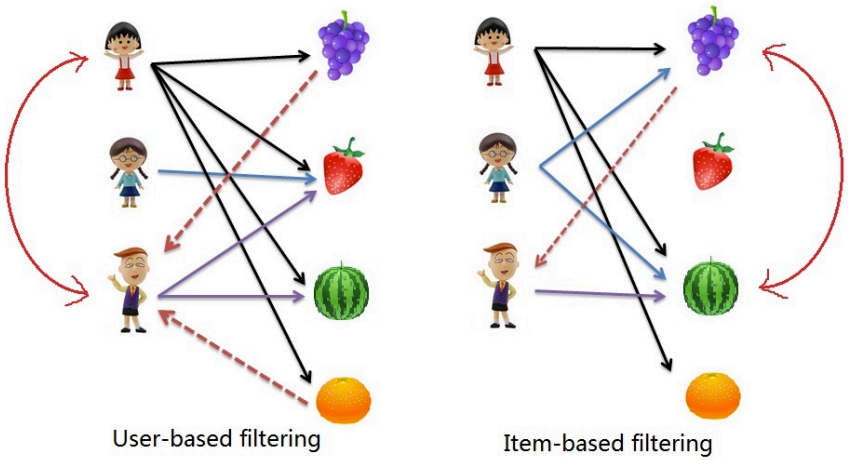


Representation learning



Recommender Systems

- User-based collaborative filtering
- Item-based collaborative filtering



- Sparse matrix completion
  - ❑Netflix problem

Graphical Models: Topic Modeling