Natural Language Processing
SoSe 2015

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Machine Learning for NLP

Dr. Mariana Neves

May 4th, 2015

(based on the slides of Dr. Saeedeh Momtazi)

Introduction

- „Field of study that gives computers the ability to learn without being explicitly programmed"
  - Arthur Samuel, 1959

(http://www.cs.toronto.edu/~urtasun/courses/CSC2515/CSC2515_Winter15.html)

Introduction

- Learning Methods
  - Supervised learning
    - Active learning
  - Unsupervised learning
  - Semi-supervised learning
  - Reinforcement learning

Outline

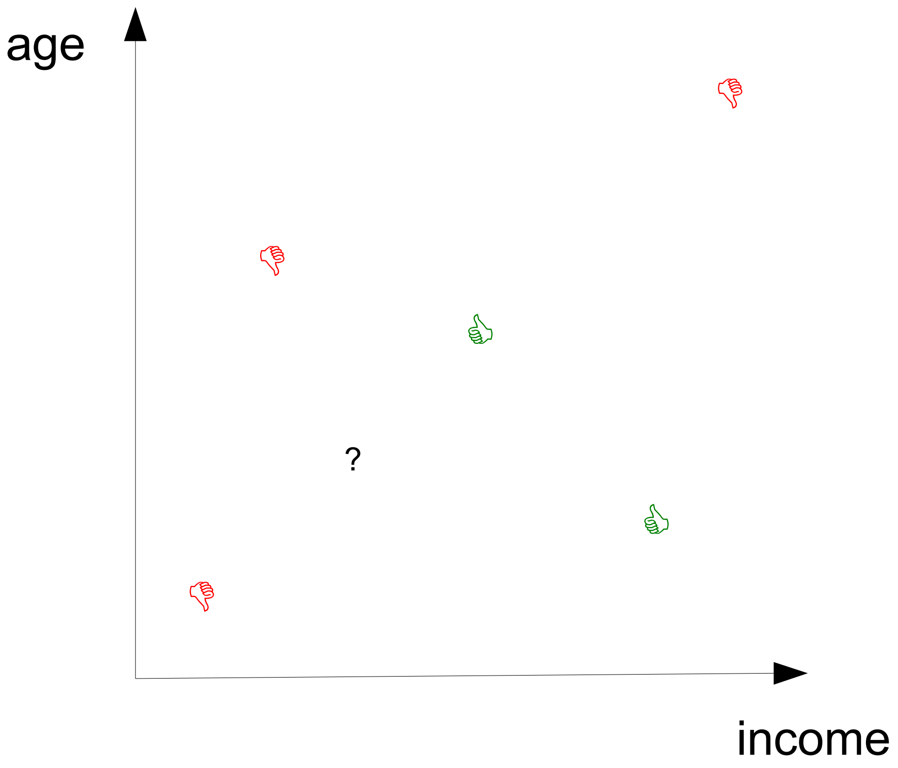
- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

Supervised Learning

- Example: mortgage credit decision
  - Age
  - Income

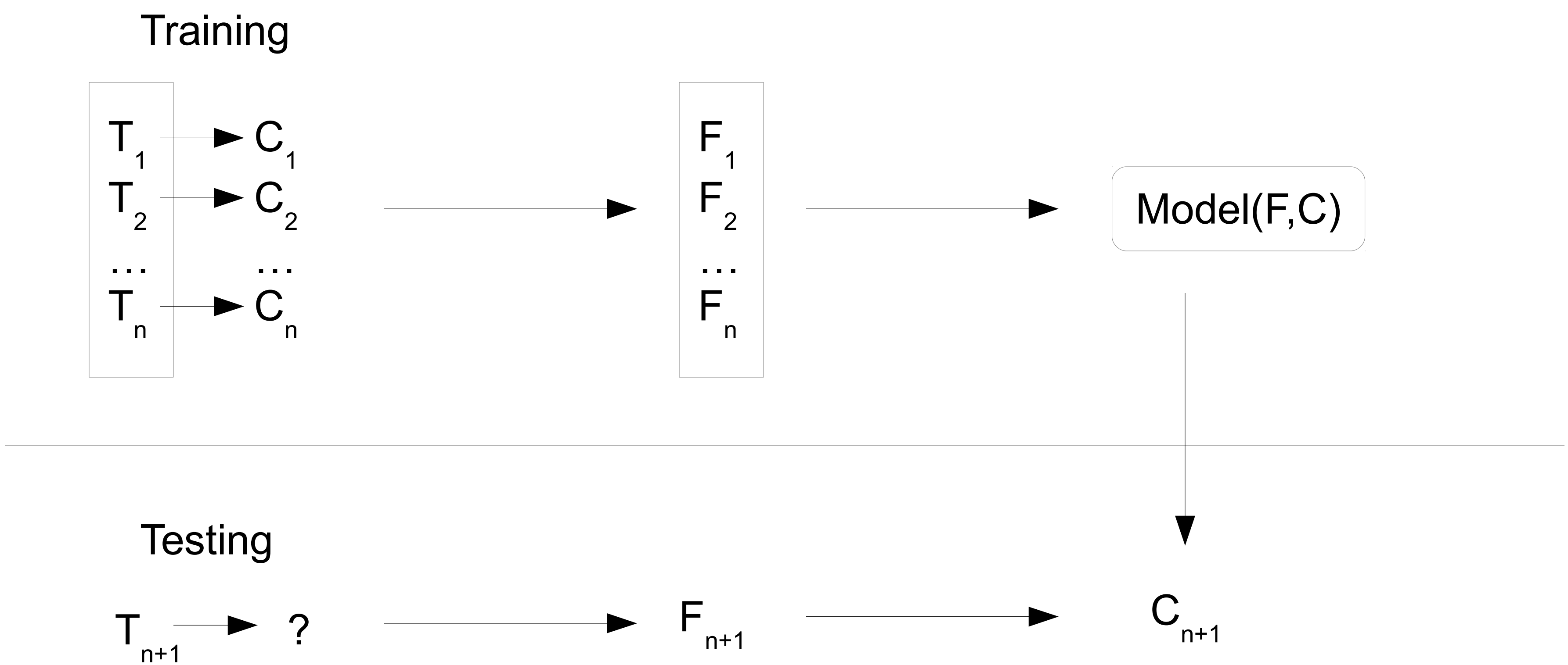
http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules

Supervised Learning



http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules

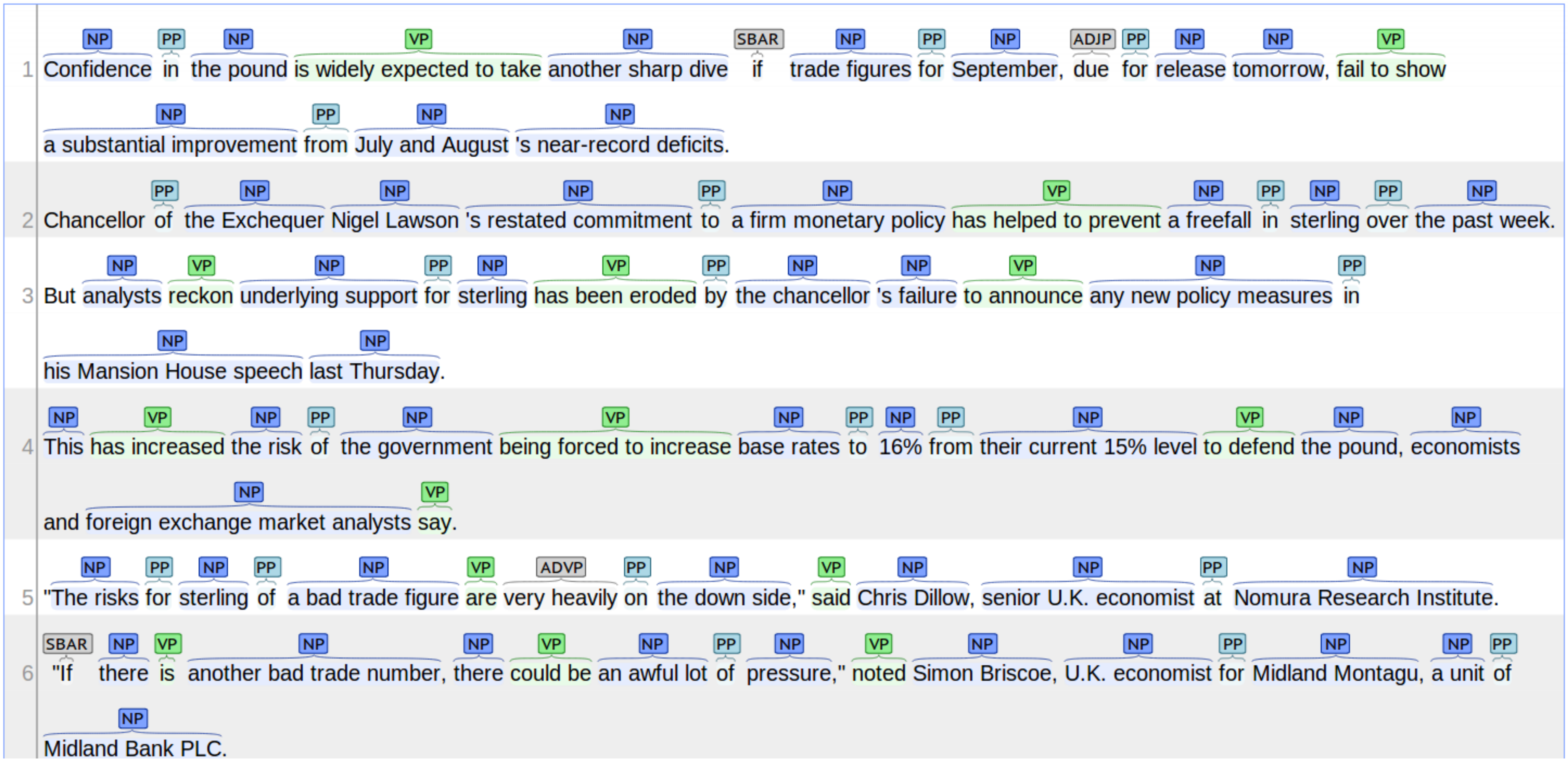
Classification



Applications

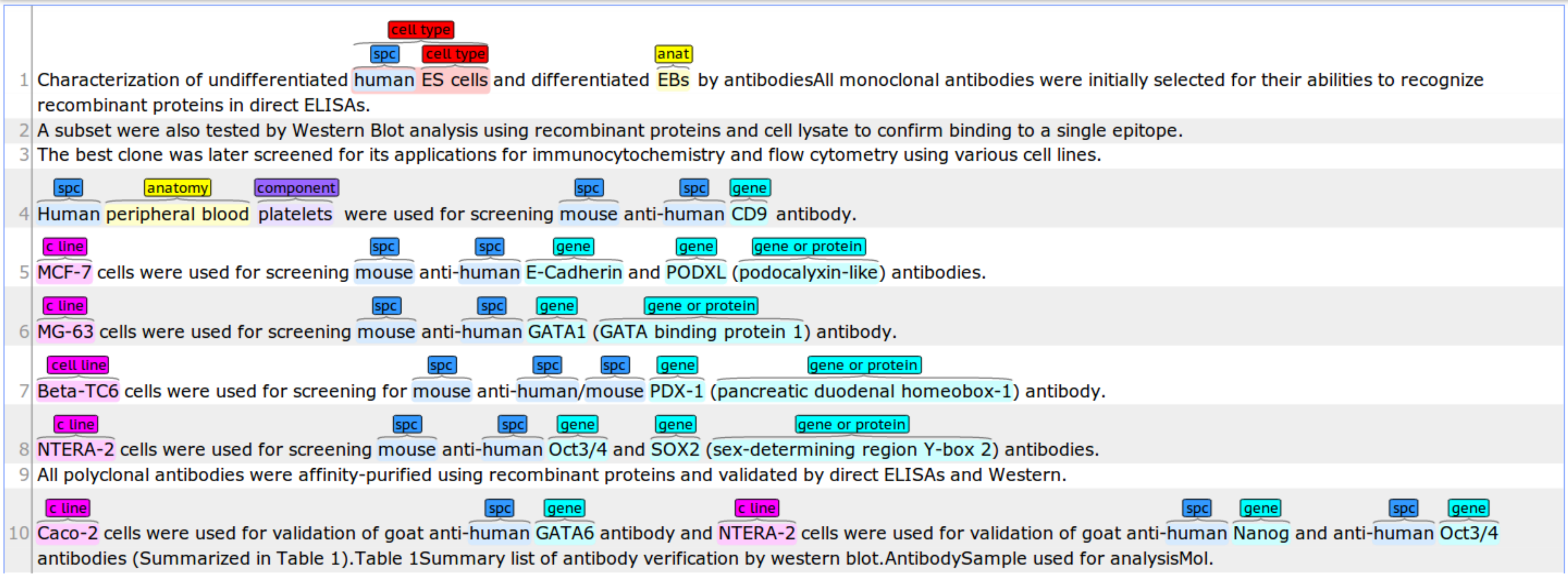
Problems	Items	Categories
POS tagging	Word	POS
Named entity recognition	Word	Named entity
Word sense disambiguation	Word	Word's sense
Spam mail detection	Document	Spam/Not Spam
Language identification	Document	Language
Text categorization	Document	Topic
Information retrieval	Document	Relevant/Not relevant

Part-of-speech tagging



http://weaver.nlplab.org/~brat/demo/latest/#/not-editable/CoNLL-00-Chunking/train.txt-doc-1

Named entity recognition



http://corpora.informatik.hu-berlin.de/index.xhtml#/cellfinder/version1_sections/16316465_03_results

Word sense disambiguation

Michael Jordan (born 1963) is an American basketball player.

Michael Jordan may also refer to:

- Michael Jordan (mycologist), English mycologist
- Michael Jordan (footballer) (born 1986), English goalkeeper (Arsenal, Chesterfield, Lewes)
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Mike Jordan (born 1958), English racing driver
- Mike Jordan (baseball) (1863–1940), baseball player
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1957), American researcher in machine learning and artificial intelligence
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordan (born 1990), Czech ice hockey player
- "Michael Jordan", a song by Kendrick Lamar featuring ScHoolboy Q on the album Overly Dedicated'

Spam mail detection

Neue Nachricht
Peter Schmidt [noreply@comment.am]

Sent: Tuesday, April 29, 2014 10:32 AM
To: Forschungskolleg

Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen bietet, m?ssten Sie einfach nur kurz einen kostenlosen Vergleich auf unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich. Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

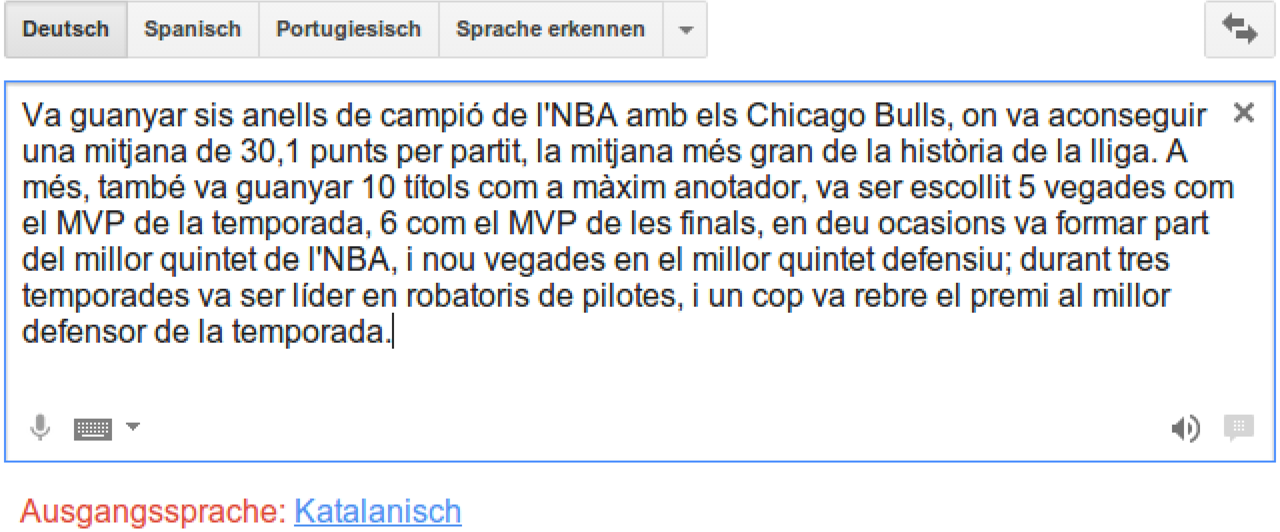
Besuchen Sie unsere Webseite unter:

http://www.pkv-check2014.com

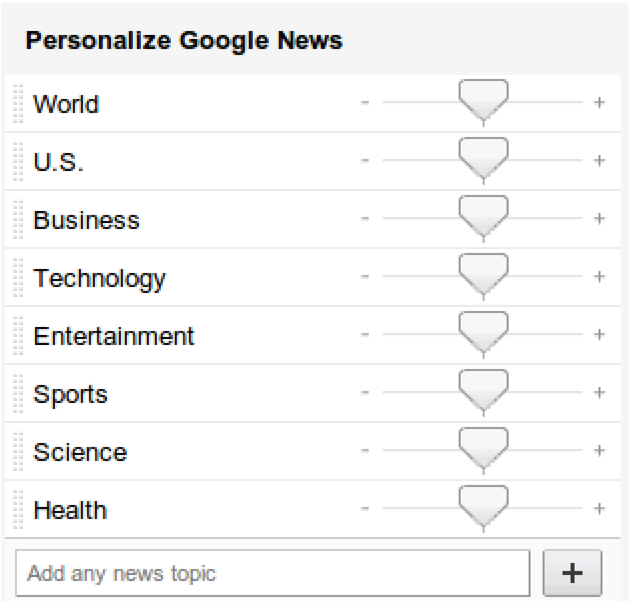
Ich hoffe ich konnte Ihnen helfen

Aus Newsletter austragen unter:
http://www.pkv-check2014.com/unsubscribe

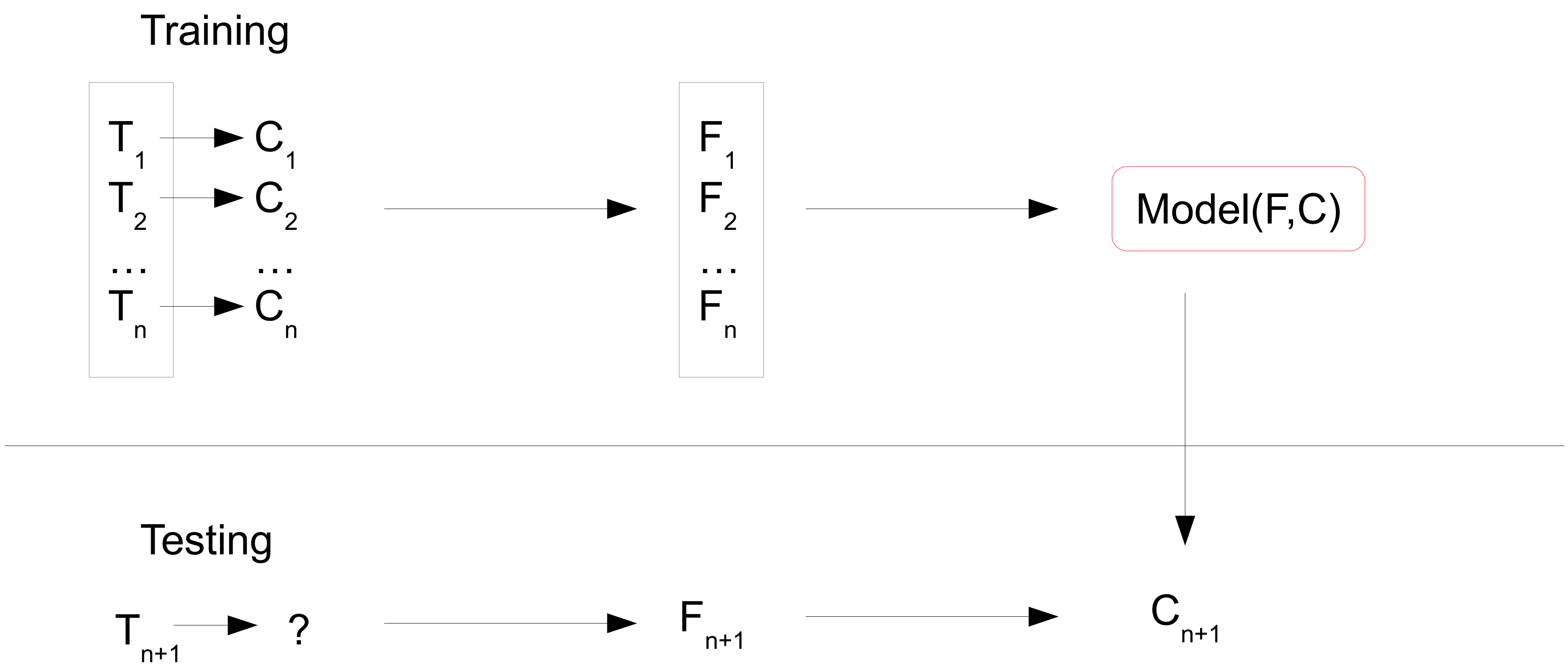
Language identification



Text categorization



Classification



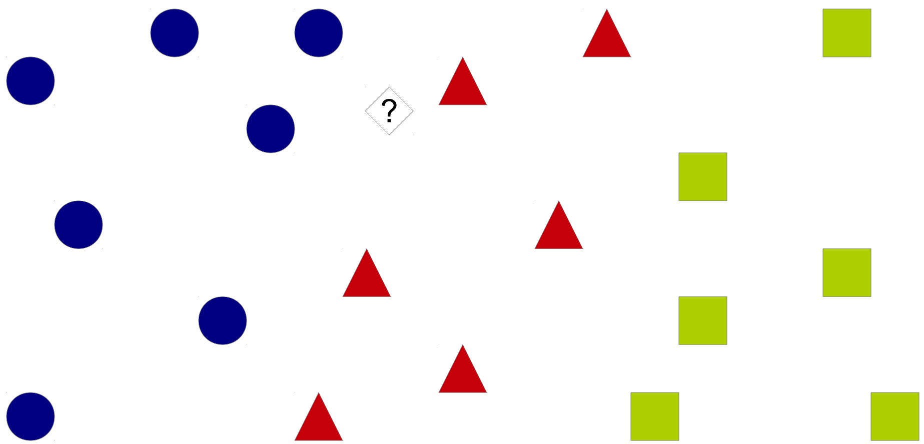
Classification algorithms

- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

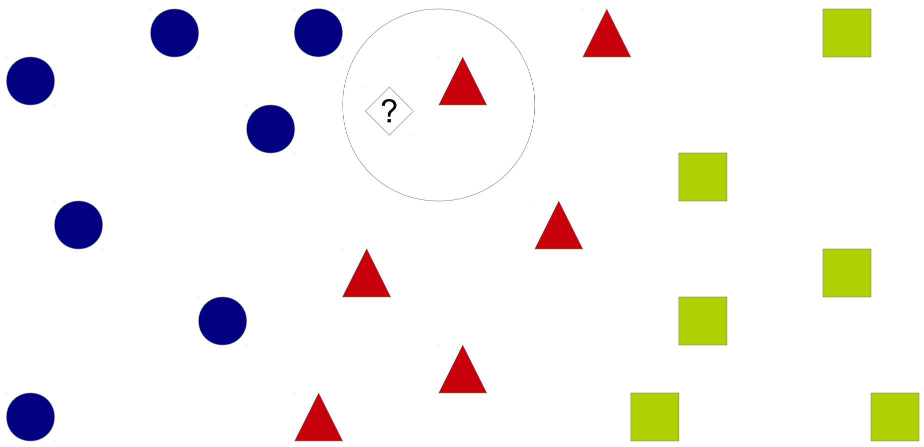
Classification algorithms

- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

K Nearest Neighbor

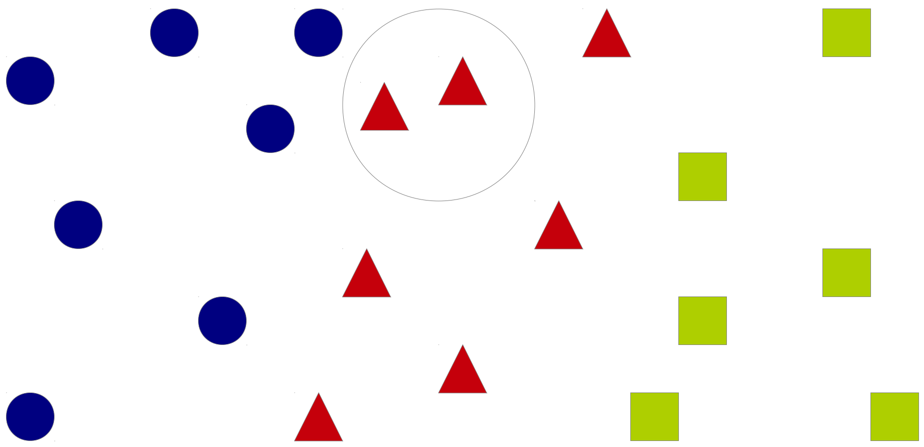


K Nearest Neighbor



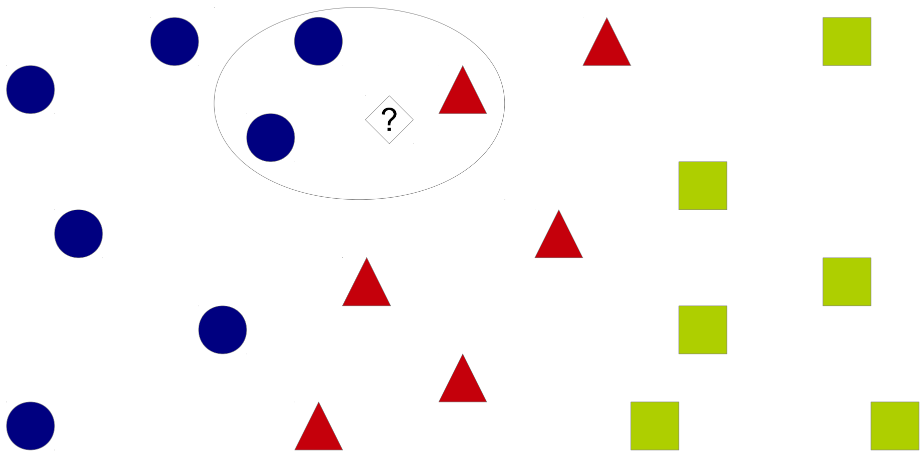
K Nearest Neighbor

- 1-nearest neighbor



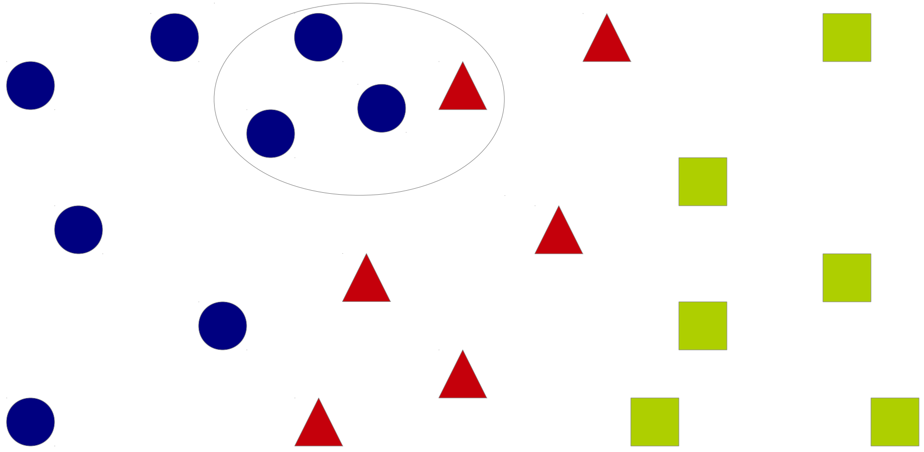
K Nearest Neighbor

- 3-nearest neighbors



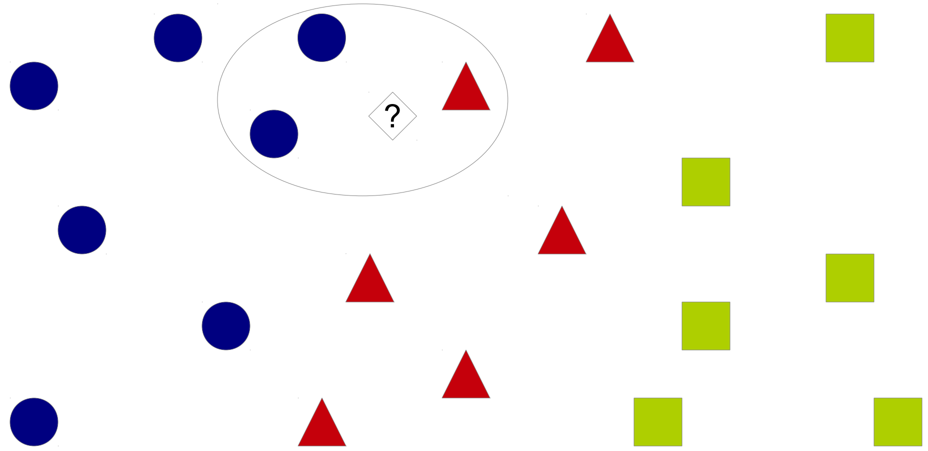
K Nearest Neighbor

- 3-nearest neighbors



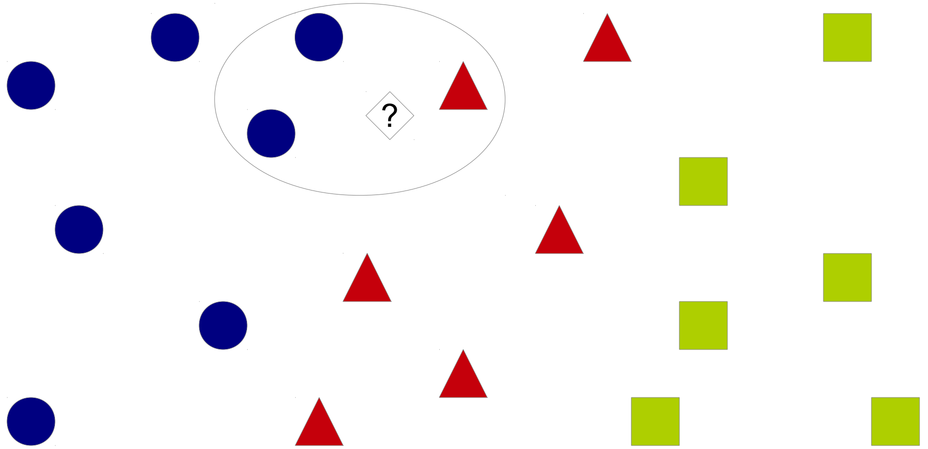
K Nearest Neighbor

- Simple approach
- No black-box
- Choose
  - Features
  - Distance metrics
  - Value of k (majority voting)



K Nearest Neighbor

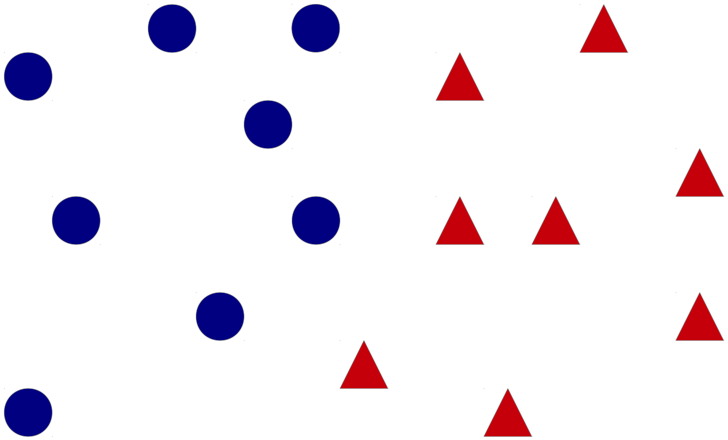
- Class distribution is skewed



Classification algorithms

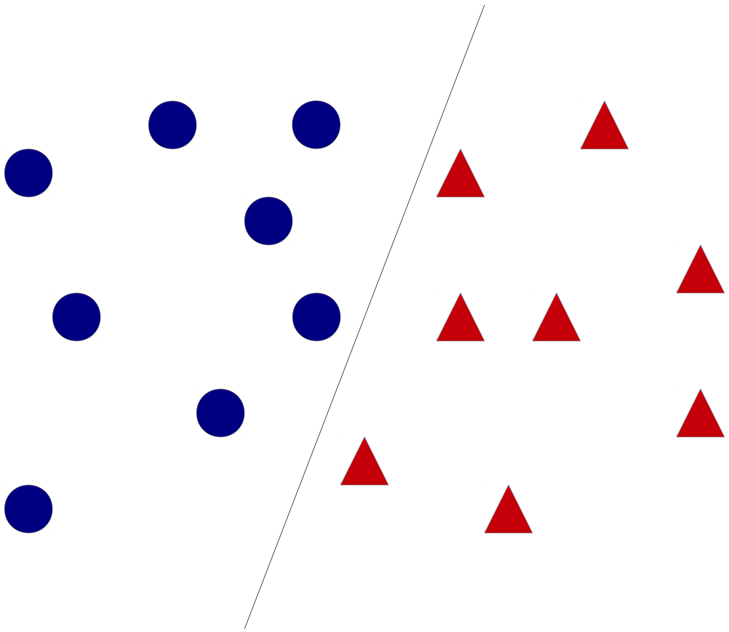
- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

Support vector machines



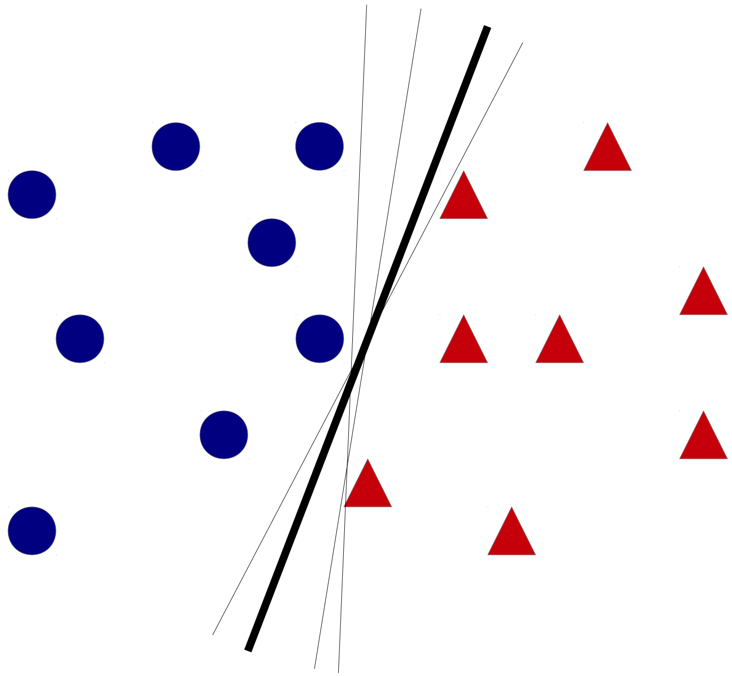
Support vector machines

- Find a hyperplane in the vector space that separates the items of the two categories



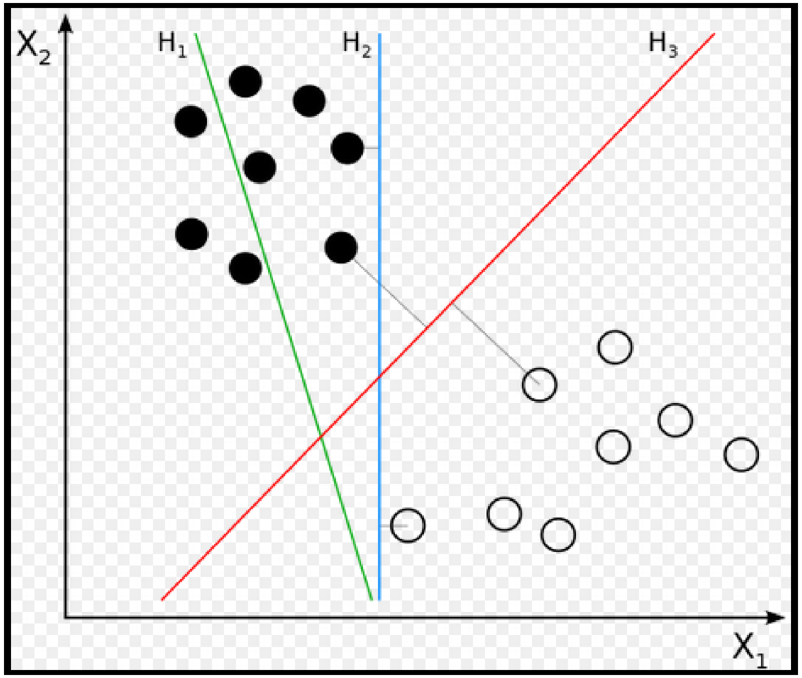
Support vector machines

- There might be more than one possible separating hyperplane



Support vector machines

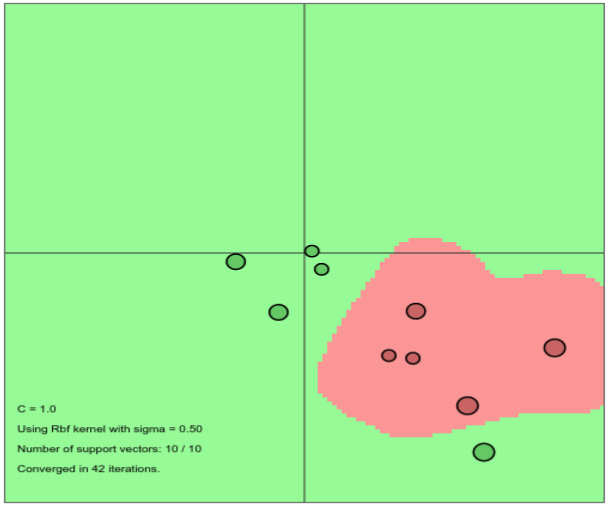
- Find the hyperplane with maximum margin
- Vectors at the margins are called support vectors



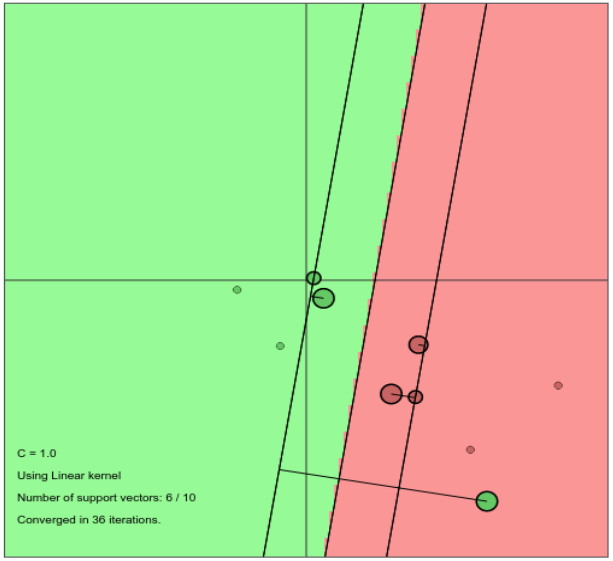
(http://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Svm_separating_hyperplanes_(SVG).svg)

Support vector machines

- Usually provides good results



- There are libraries available for many programming languages



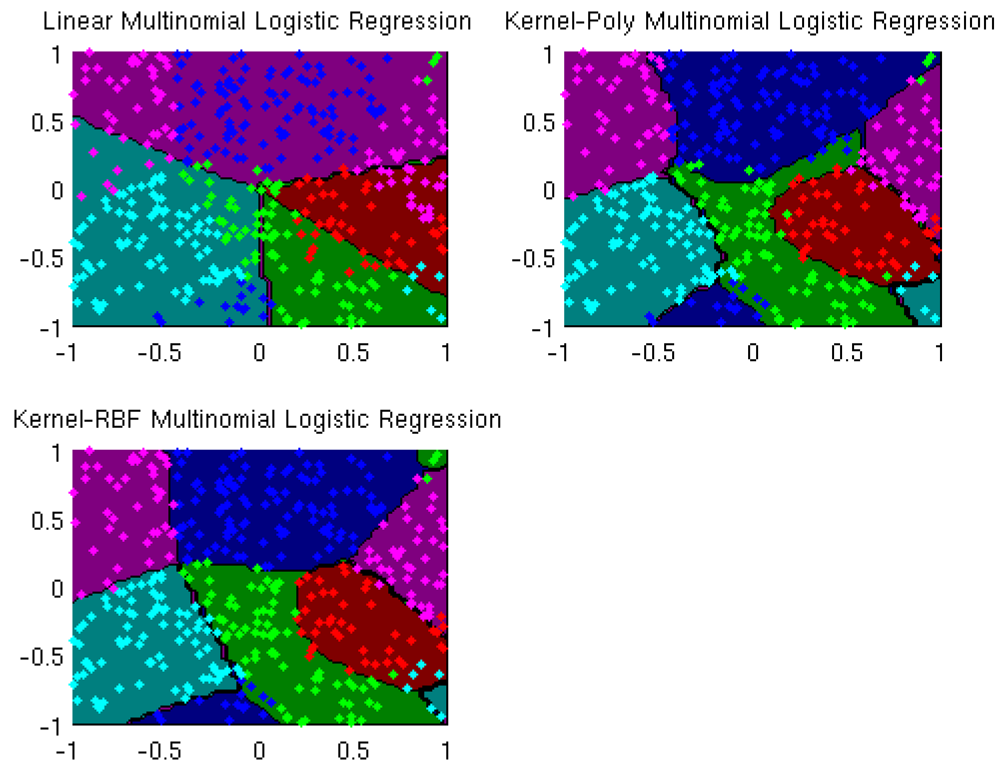
- Linear and non-linear classification

(http://cs.stanford.edu/people/karpathy/svmjs/demo/)

Support vector machines

- Black-box

- Multi-class problems usually modelled as many binary classifications



(http://www.cs.ubc.ca/~schmidtm/Software/minFunc/examples.html)

Classification algorithms

- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

Naïve Bayes

- Selecting the class with highest probability
  - Minimizing the number of items with wrong labels

$$\hat{c} = argmax_{c_i} P(c_i)$$

  - Probability should depend on the to be classified data (d)

$$P(c_i|d)$$

Naïve Bayes

$$\hat{c} = argmax_{c_i} P(c_i)$$

$$\hat{c} = argmax_{c_i} P(c_i|d)$$

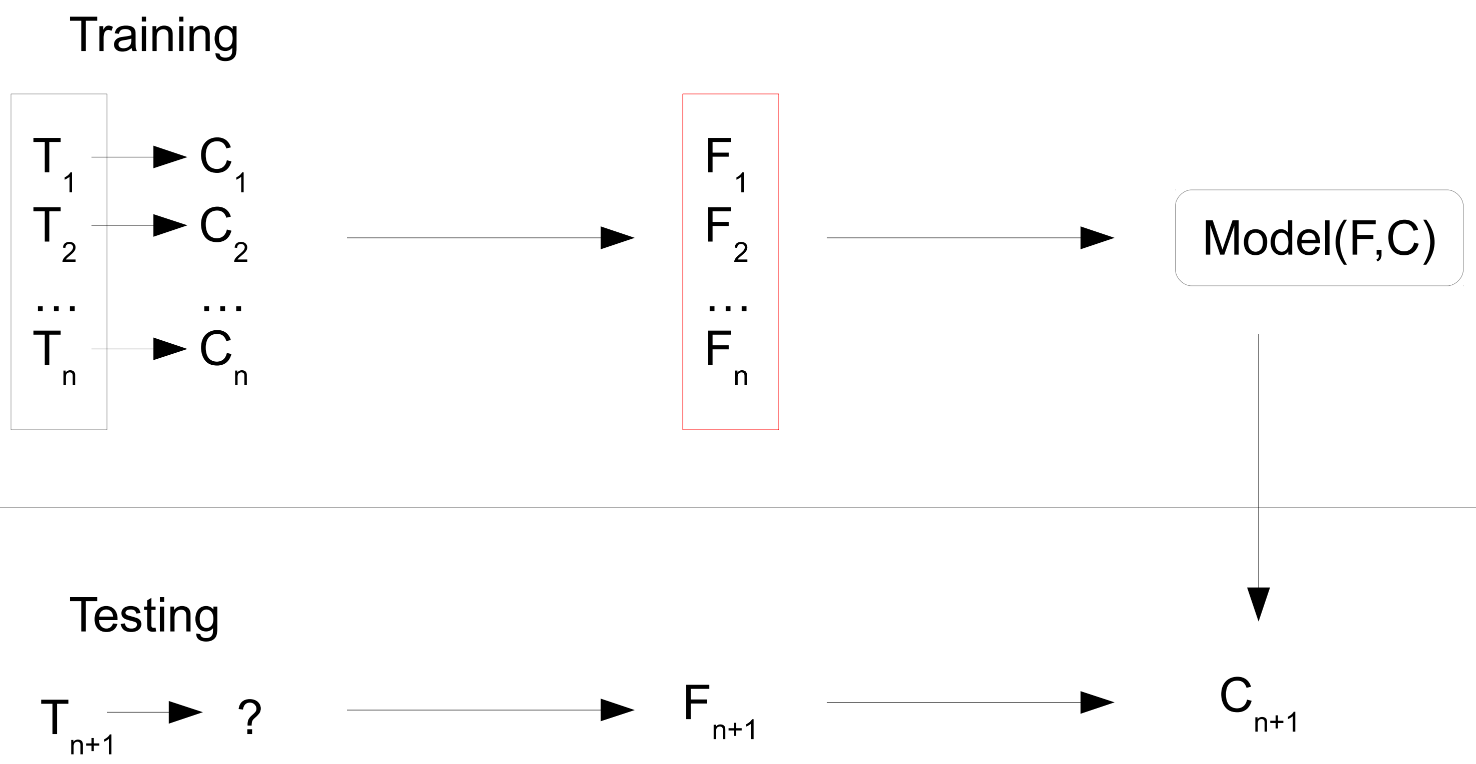
$$\hat{c} = argmax_{c_i} \frac{P(d|c_i) \cdot P(c_i)}{P(d)}$$

$$\hat{c} = argmax_{c_i} P(d|c_i) \cdot P(c_i)$$

Naïve Bayes

$$\hat{c} = argmax_{c_i} \underbrace{P(d|c_i)}_{\text{Likelihood probability}} \cdot \underbrace{P(c_i)}_{\text{Prior probability}}$$

Classification



Spam mail detection

Neue Nachricht
Peter Schmidt [noreply@comment.am]

Sent: Tuesday, April 29, 2014 10:32 AM
To: Forschungskolleg

Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen bietet, m?ssten Sie einfach nur kurz einen kostenlosen Vergleich auf unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich. Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

Besuchen Sie unsere Webseite unter:

http://www.pkv-check2014.com

Ich hoffe ich konnte Ihnen helfen

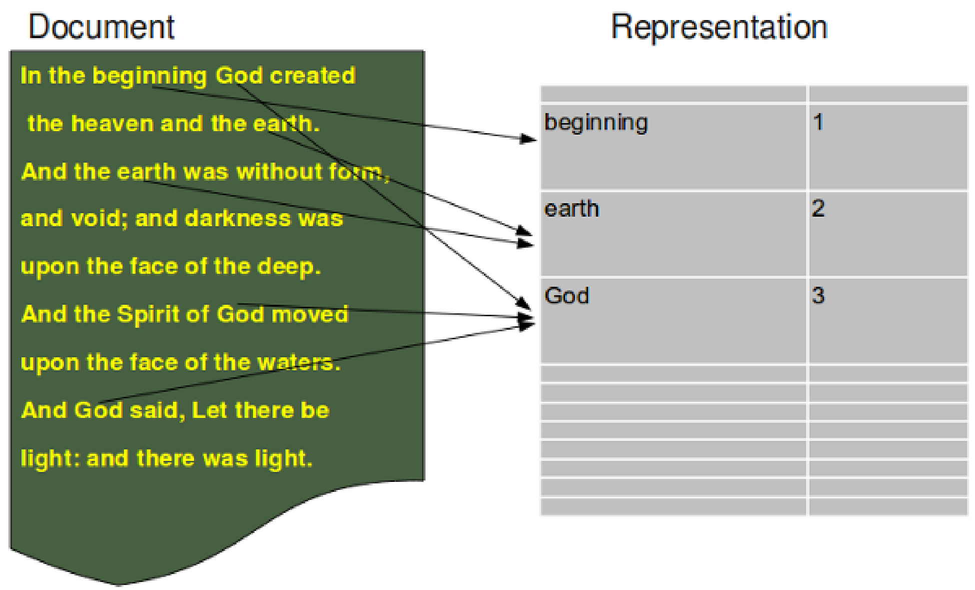
Aus Newsletter austragen unter:
http://www.pkv-check2014.com/unsubscribe

Features:
- words
- sender's email
- contains links
- contains attachments
- contains money amounts

...

Feature selection

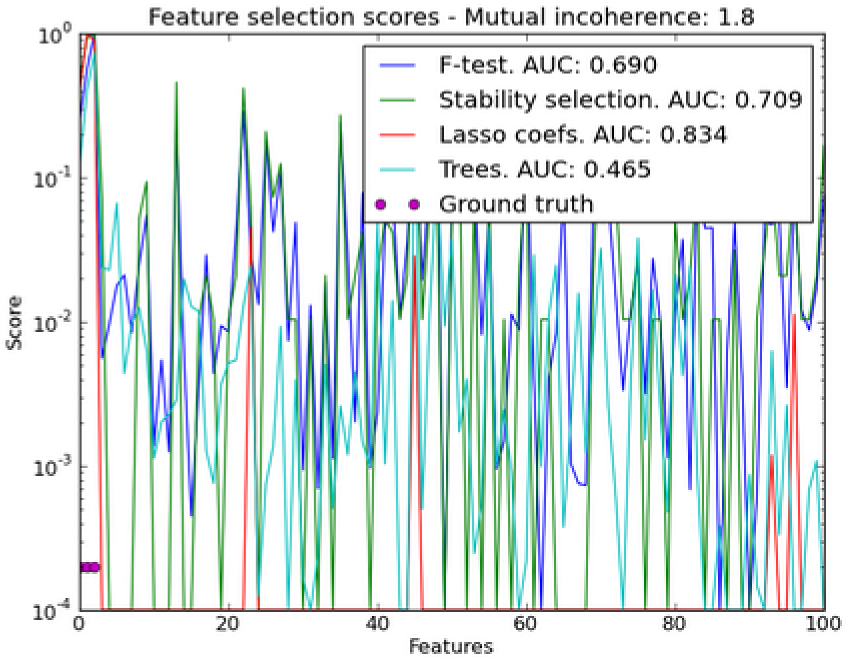
- Bag-of-words:
  - Document represented by the set of words
  - High dimensional feature space
  - Computationally expensive



(http://www.python-course.eu/text_classification_python.php)

Feature selection

- Solution: Feature selection to select informative words



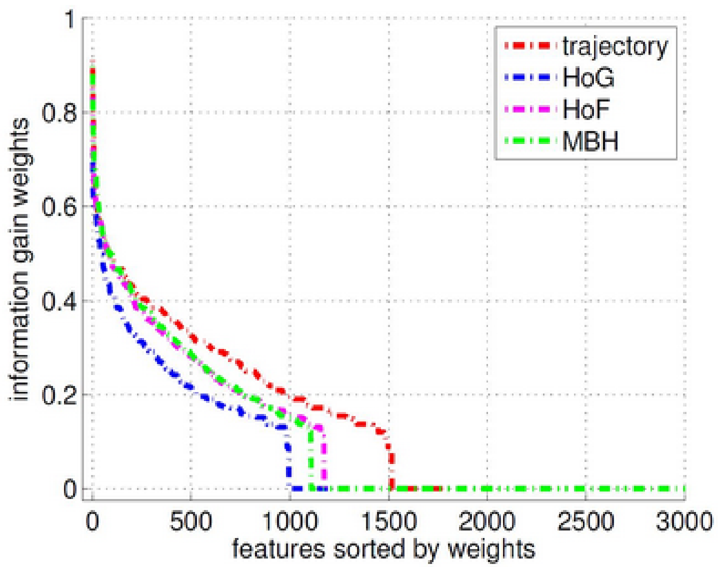
(http://scikit-learn.org/0.11/modules/feature_selection.html)

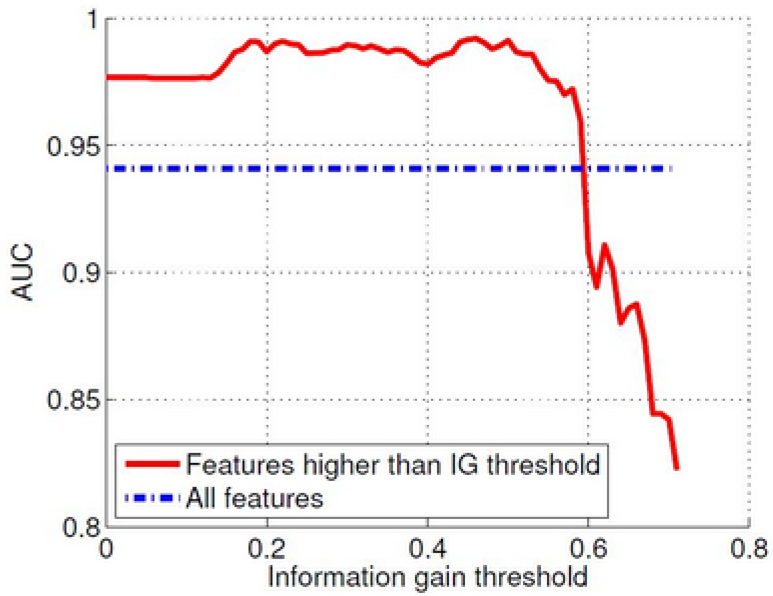
Feature selection methods

- Information gain
- Mutual information
- χ-Square

Information gain

- Measuring the presence or absence of a term in the document
- Removing words whose information gain is less than a predefined threshold





(http://seqam.rutgers.edu/site/index.php?option=com_jresearch&view=project&task=show&id=35)

Information gain

$$IG(w) = \sum_{i=1}^{K} P(c_i) \cdot \log P(c_i)$$

$$+ P(w) \cdot \sum_{i=1}^{K} P(c_i | w) \cdot \log P(c_i | w)$$

$$+ P(\bar{w}) \cdot \sum_{i=1}^{K} P(c_i | \bar{w}) \cdot \log P(c_i | \bar{w})$$

Information gain

- N = # docs
- N_i = # docs in category c_i
- N_w = # docs containing w
- $N_{\bar{w}}$ = # docs not containing w
- N_{iw} = # docs in category c_i containing w
- $N_{i\bar{w}}$ = *# docs in category c_i not containing w*

$$P(c_i) = \frac{N_i}{N}$$

$$P(w) = \frac{N_w}{N} \qquad P(c_i|w) = \frac{N_{iw}}{N_i}$$

$$P(\bar{w}) = \frac{N_{\bar{w}}}{N} \qquad P(c_i|\bar{w}) = \frac{N_{i\bar{w}}}{N_i}$$

Mutual information

- Measuring the effect of each word in predicting the category

$$MI(w, c_i) = \log \frac{P(w, c_i)}{P(w) \cdot P(c_i)}$$

Mutual information

- Removing words whose mutual information is less than a predefined threshold

$$MI(w) = max_i \, MI(w, c_i)$$

$$MI(w) = {}_i\sum P(c_i) \cdot MI(w, c_i)$$

χ-Square

- Measuring the dependencies between words and categories

$$\chi^2(w,c_i)=\frac{N\cdot\left(N_{iw}N_{\bar{i}\bar{w}}-N_{i\bar{w}}N_{\bar{i}w}\right)^2}{\left(N_{iw}+N_{i\bar{w}}\right)\cdot\left(N_{\bar{i}w}+N_{\bar{i}\bar{w}}\right)\cdot\left(N_{iw}+N_{\bar{i}w}\right)\cdot\left(N_{i\bar{w}}+N_{\bar{i}\bar{w}}\right)}$$

- Ranking words based on their χ-square measure

$$\chi^2(w)=\sum_{i=1}^{K}P(c_i)\cdot\chi^2(w,c_i)$$

- Selecting the top words as features

Feature selection

- These models perform well for document-level classification
  - Spam Mail Detection
  - Language Identification
  - Text Categorization

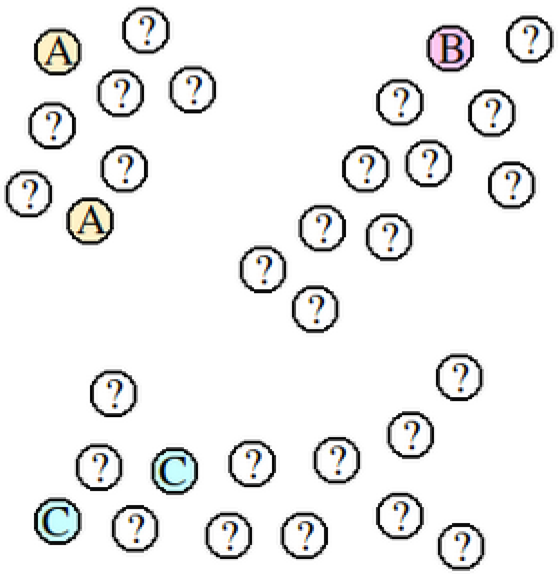
- Word-level Classification might need another types of features
  - Part-of-speech tagging
  - Named Entity Recognition

Supervised learning

- Shortcoming
  - Relies heavily on annotated data
  - Time consuming and expensive task

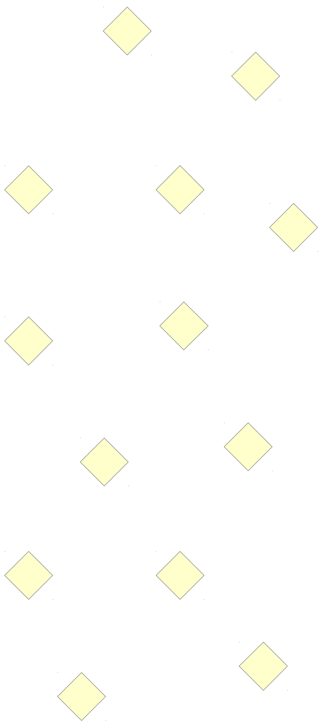
Supervised learning

- Active learning
  - Using a minimum amount of annotated data
  - Annotating further data by human, if they are very informative

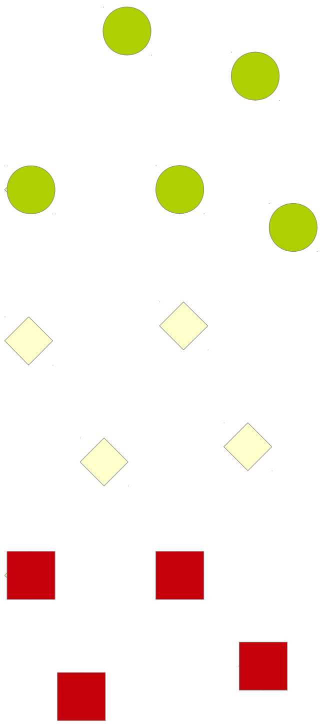


http://en.wikipedia.org/wiki/Transduction_(machine_learning)

Active learning

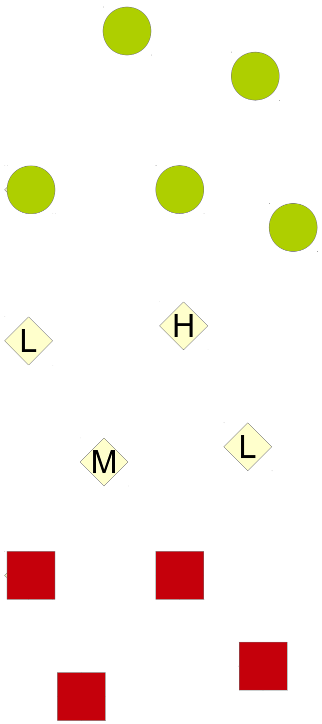


Active learning



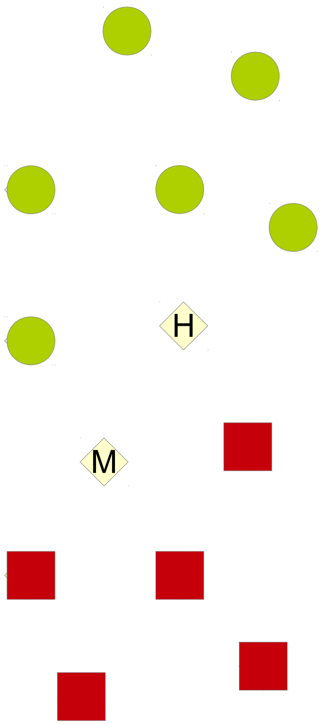
- Annotating a small amount of data

Active learning



- Calculating the confidence score of the classifier on unlabeled data

Active learning



- Finding the informative unlabeled data (data with lowest confidence)

- manually annotating the informative data

Outline

- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

Semi-supervised learning

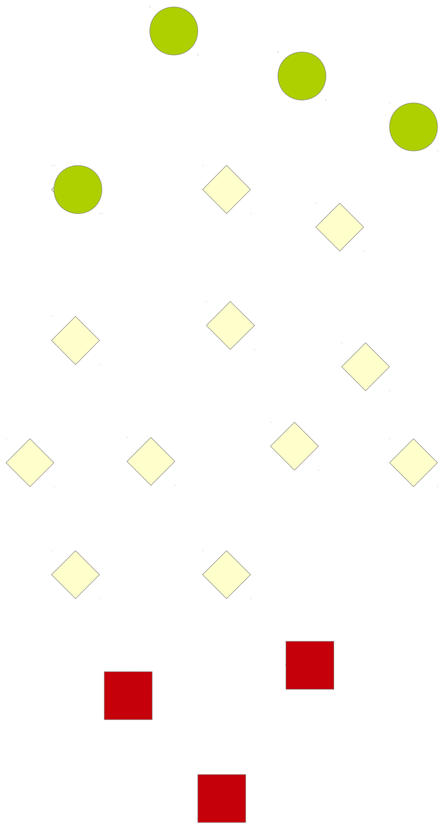
- Annotating data is a time consuming and expensive task

- Solution
  - Using a minimum amount of annotated data
  - Annotating further data automatically

Semi-supervised learning

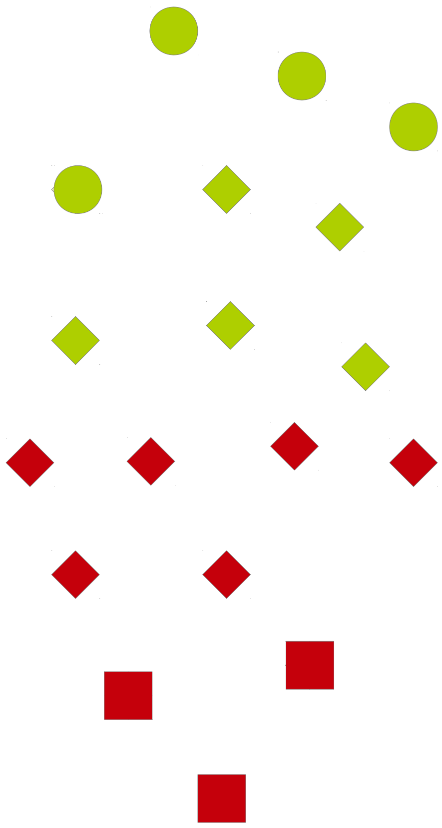
- A small amount of labeled data

Semi-supervised learning



- A large amount of unlabeled data

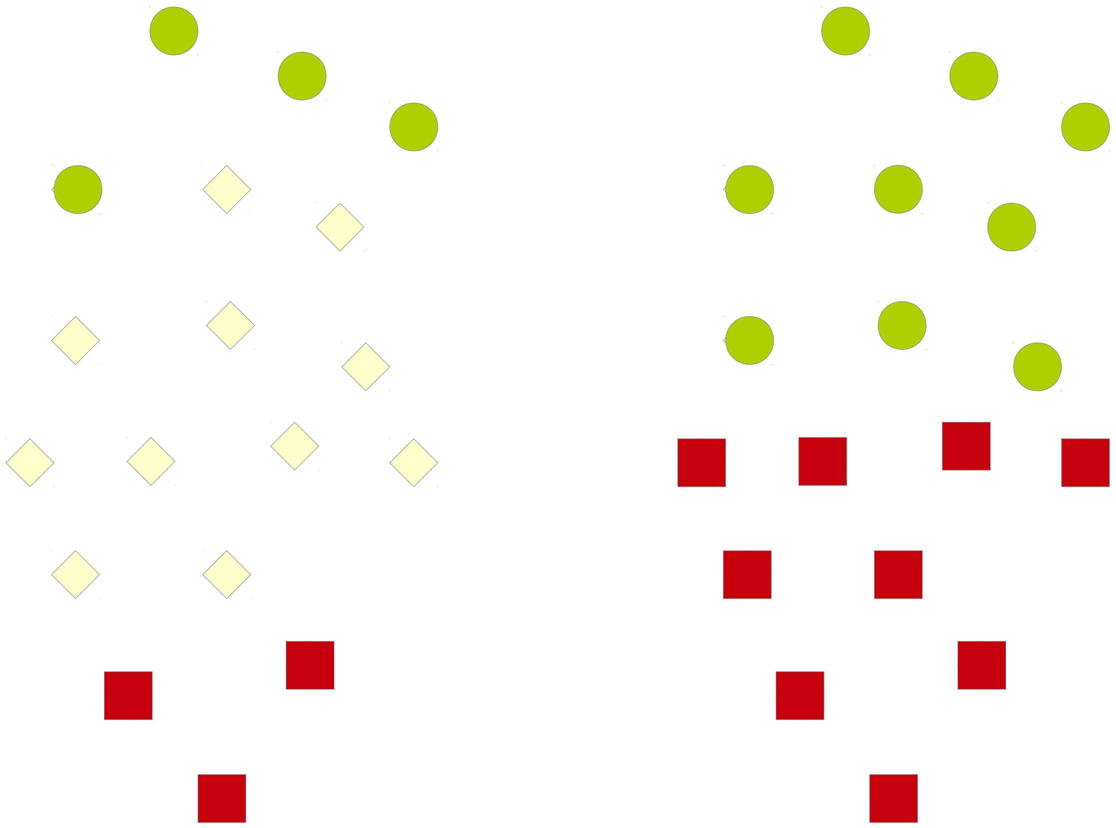
Semi-supervised learning



- Similarity between the labeled and unlabeled data

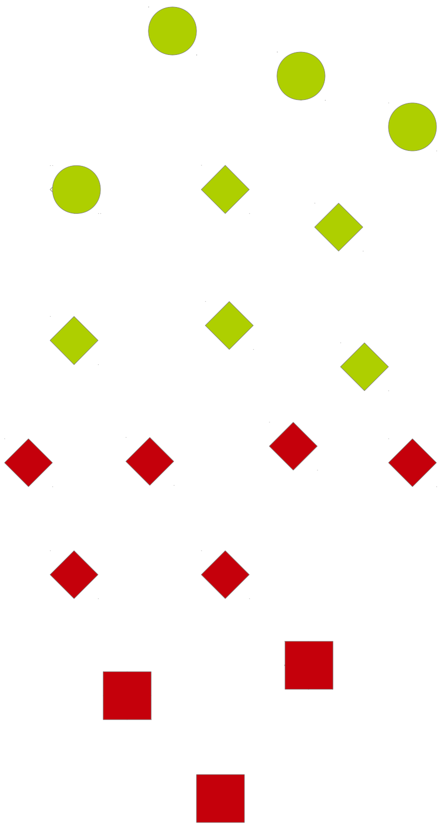
- Predicting the labels of the unlabeled data

Semi-supervised learning



- Training the classifier using labeled data and predicted labels of unlabeled data

Semi-supervised learning



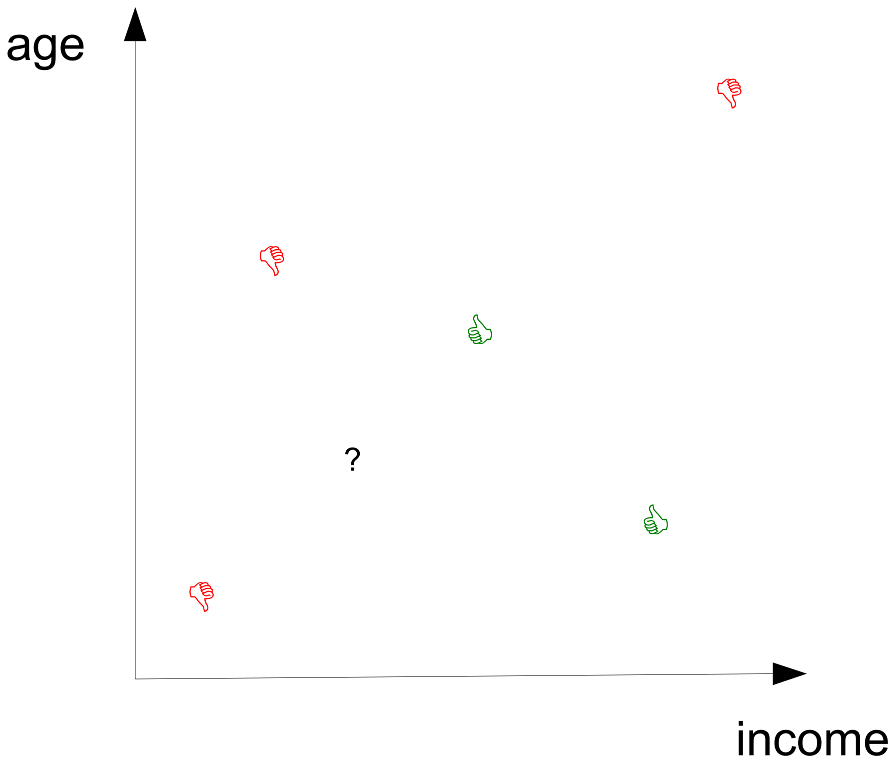
- Introduce noisy data to the system

- Add only predicted label which has high confidence

Outline

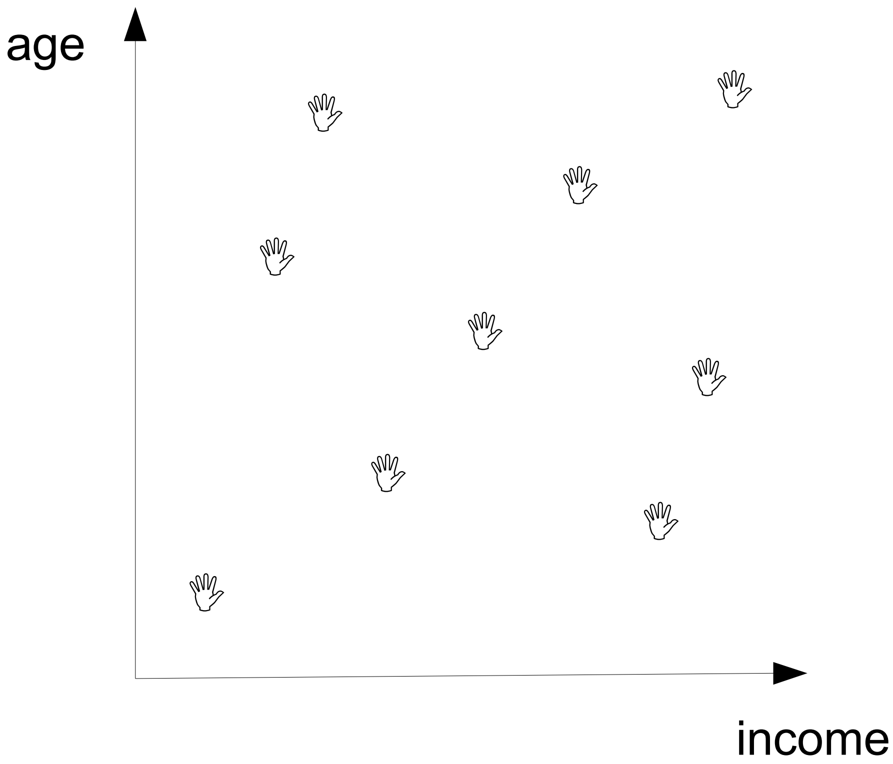
- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

Supervised Learning



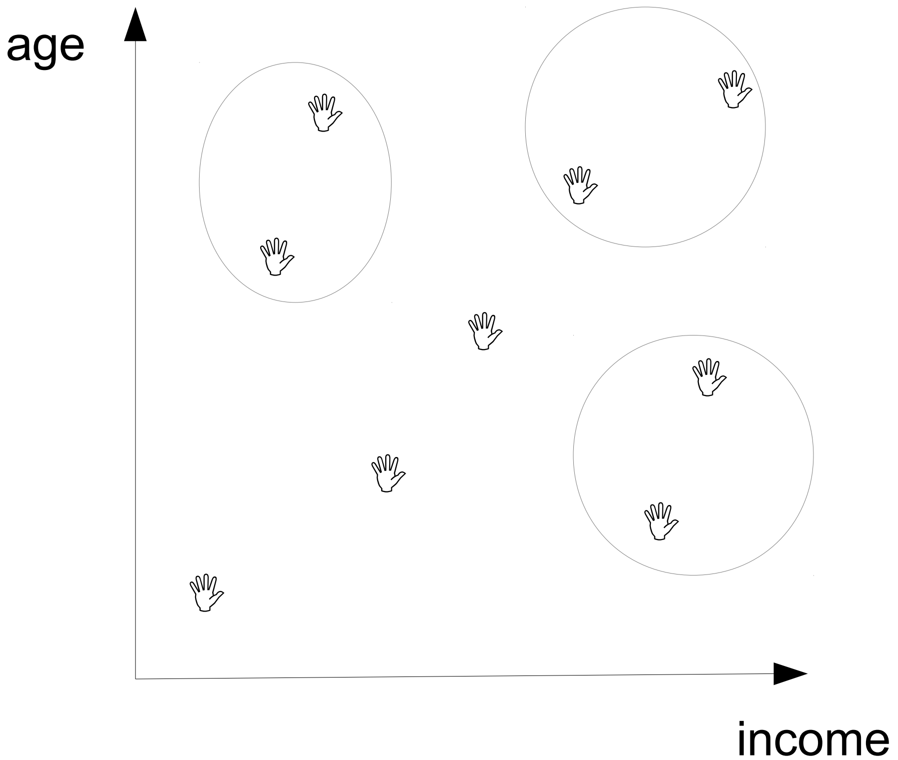
http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules

Unsupervised Learning



http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules

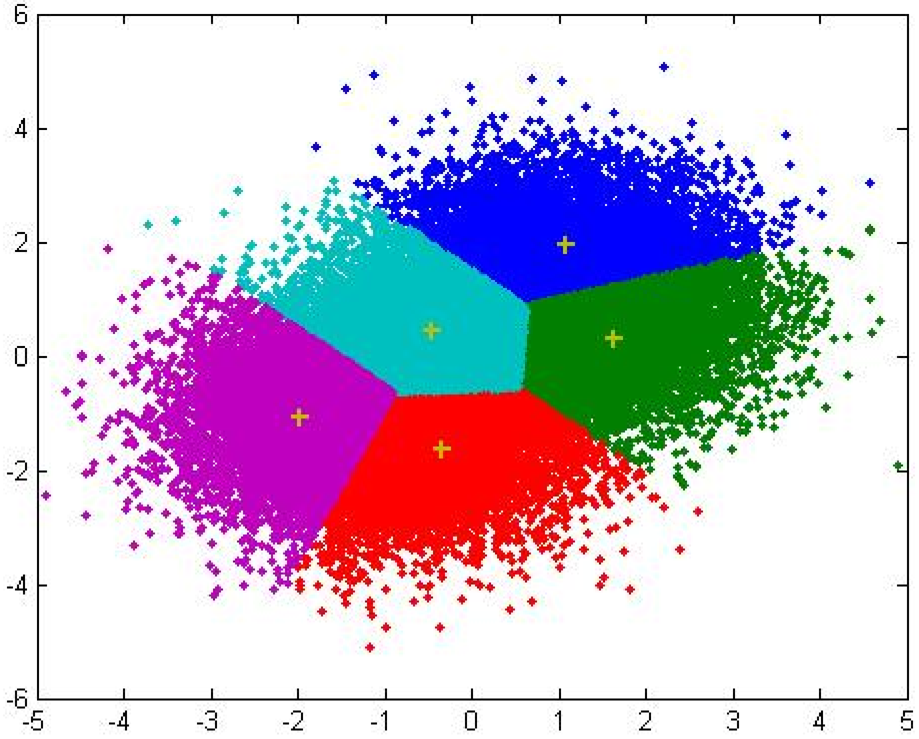
Unsupervised Learning



http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules

Clustering

- Calculating similarities between the data items
- Grouping similar data items to the same cluster



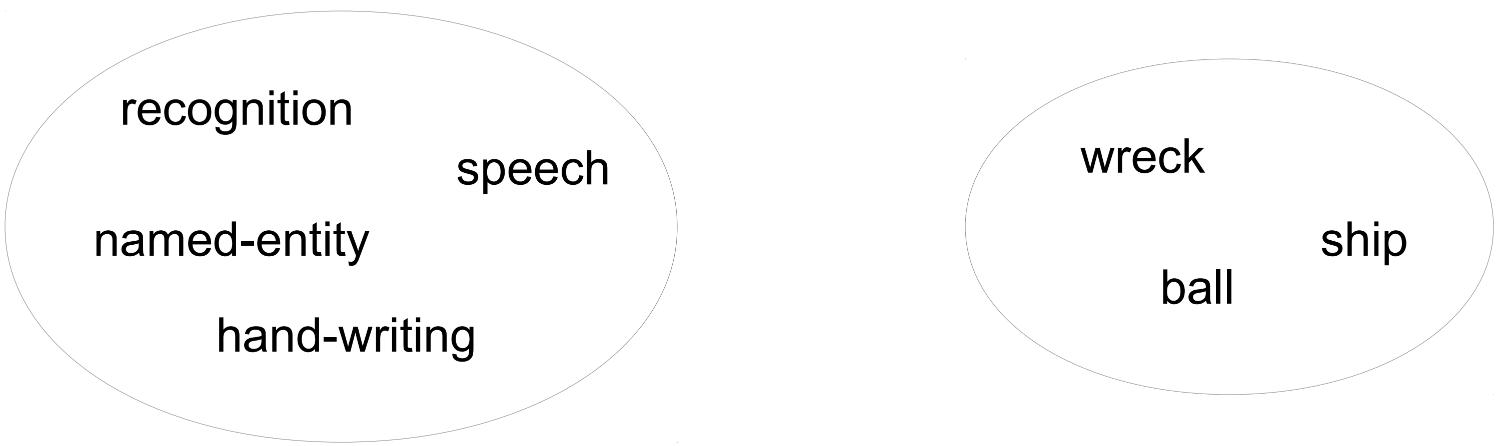
Applications

- Word clustering
  - Speech recognition
  - Machine translation
  - Named entity recognition
  - Information retrieval
  - ...

- Document clustering
  - Text classification
  - Information retrieval
  - ...

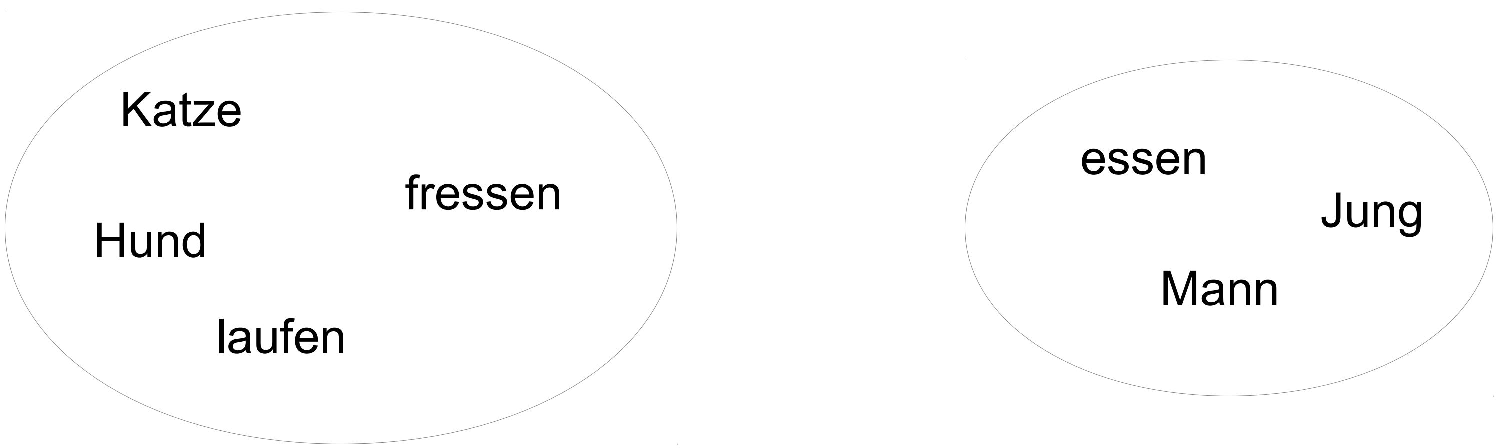
Speech recognition

- „Computers can recognize a speeech.“
- „Computers can wreck a nice peach.“



Machine translation

- „The cat eats...“
  - „Die Katze frisst...“
  - „Die Katze isst...“



Clustering algorithms

- Flat
  - K-means

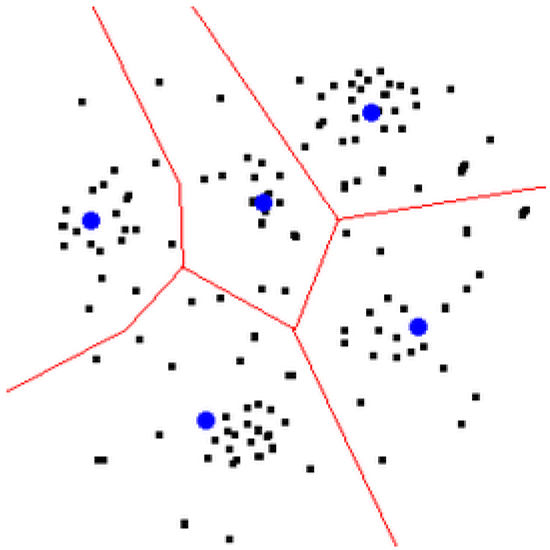
- Hierarchical
  - Top-Down (Divisive)
  - Bottom-Up (Agglomerative)
    - Single-link
    - Complete-link
    - Average-link

K-means

- The best known clustering algorithm (default/baseline), works well for many cases

$$\vec{\mu} = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- The cluster center is the mean or centroid of the items in the cluster



- Minimizing the average squared Euclidean distance of the items from their cluster centers

(http://mnemstudio.org/clustering-k-means-introduction.htm)

K-means

Initialization: Randomly choose k items as initial centroids

while stopping criterion has not been met **do**

    for each item **do**

        Find the nearest centroid

        Assign the item to the cluster associated with the nearest centroid

    end for

    for each cluster **do**

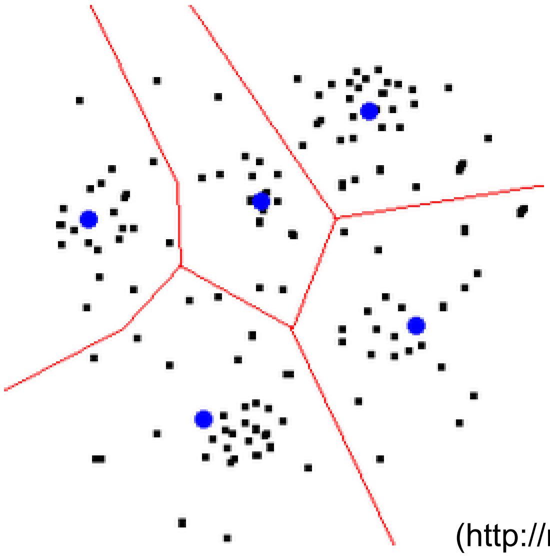
        Update the centroid of the cluster based on the average of all items in the cluster

    end for

end while

K-means

- Iterating two steps:
  - Re-assignment
    - Assigning each vector to its closest centroid
  - Re-computation
    - Computing each centroid as the average of the vectors that were assigned to it in re-assignment

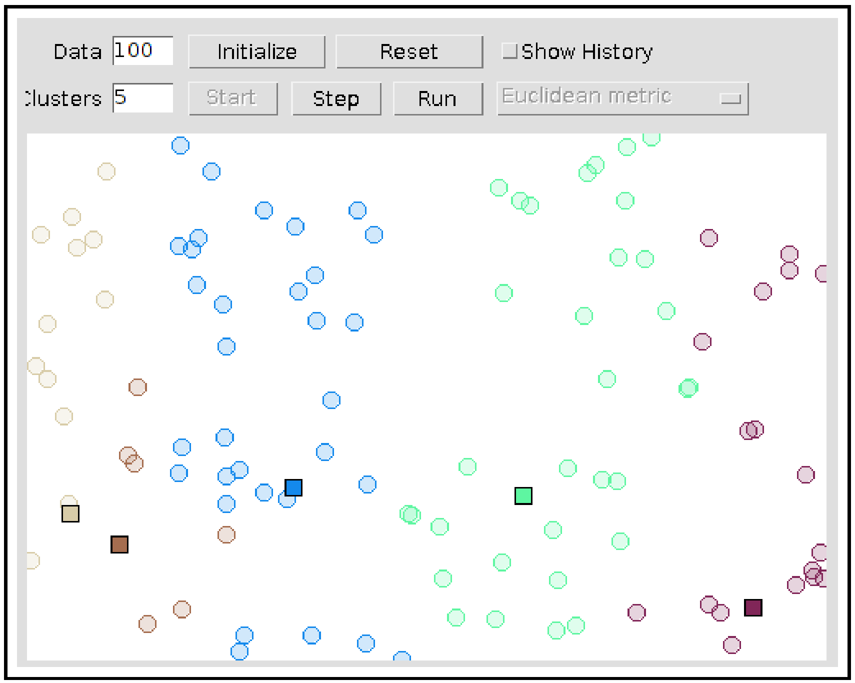


(http://mnemstudio.org/clustering-k-means-introduction.htm)

K-means

K-means - Interactive demo

This applet requires Java Runtime Environment version 1.3 or later. You can download it from the Sun Java website.

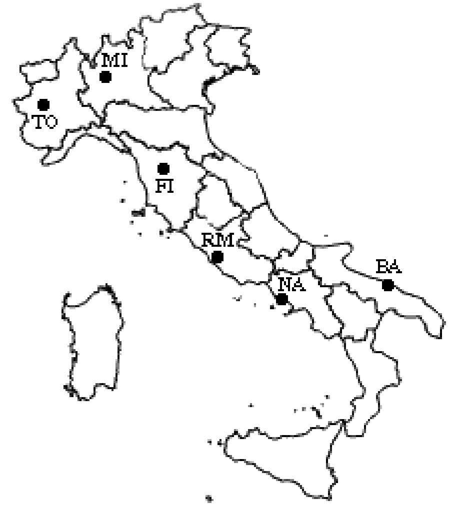


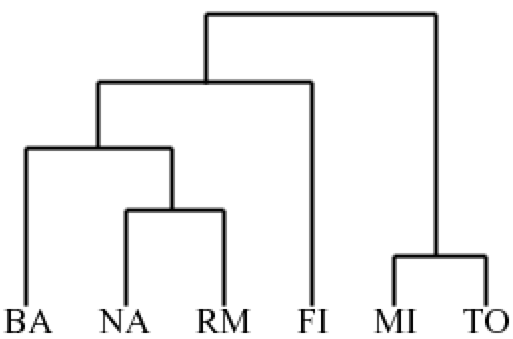
http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

Hierarchical Agglomerative Clustering (HAC)

- Creating a hierarchy in the form of a binary tree

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

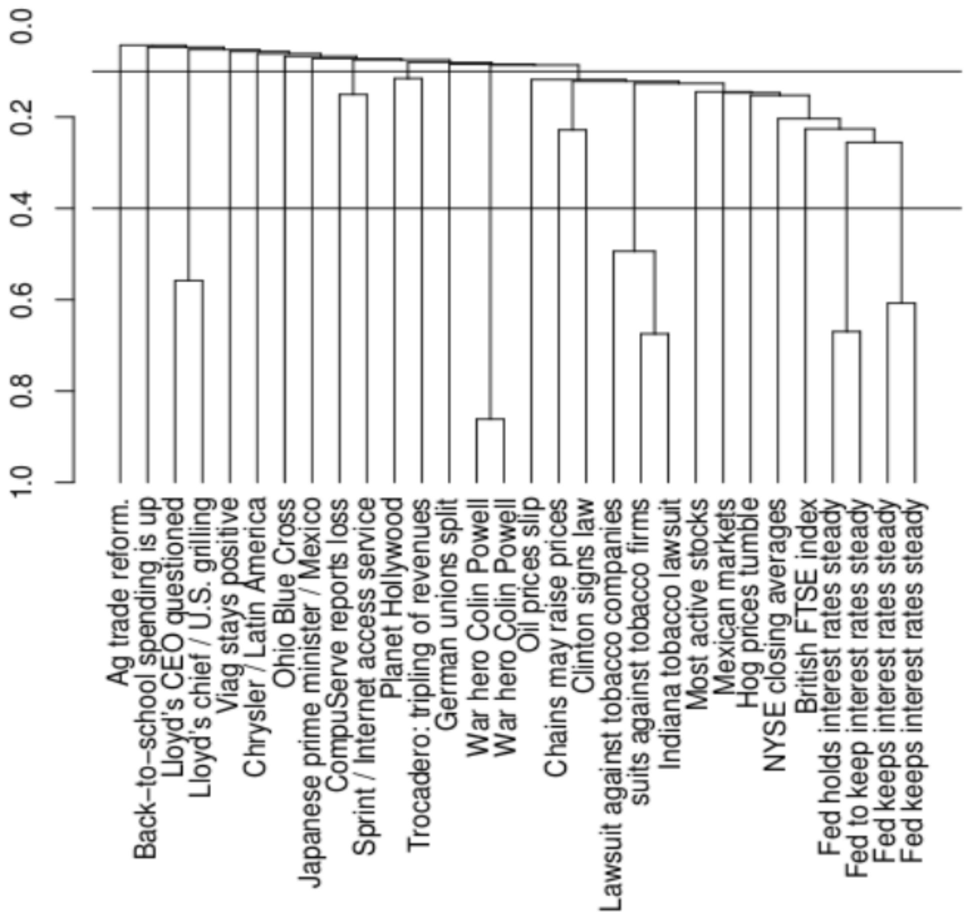




http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

Hierarchical Agglomerative Clustering (HAC)

- Creating a hierarchy in the form of a binary tree



Hierarchical Agglomerative Clustering (HAC)

Initial Mapping: Put a single item in each cluster

while reaching the predefined number of clusters **do**

    for each pair of clusters **do**

        Measure the similarity of two clusters

    end for

    Merge the two clusters that are most similar

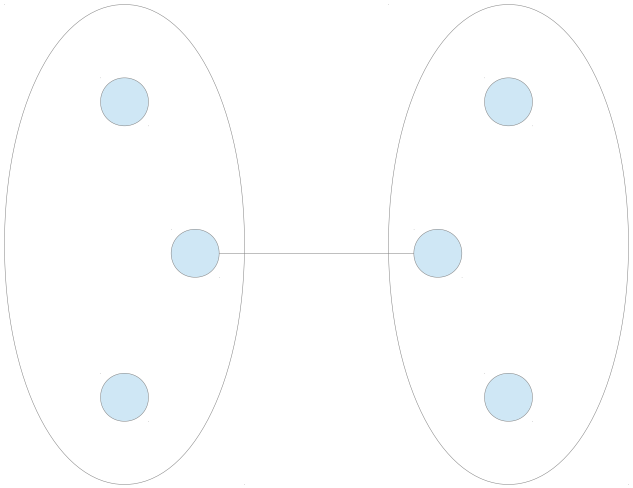
end while

Hierarchical Agglomerative Clustering (HAC)

- Measuring the similarity in three ways:
  - Single-link
  - Complete-link
  - Average-link

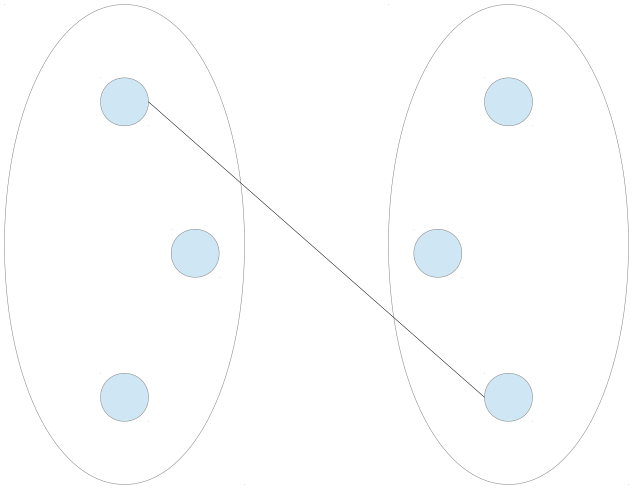
Hierarchical Agglomerative Clustering (HAC)

- Single-link / single-linkage clustering
  - Based on the similarity of the most similar members



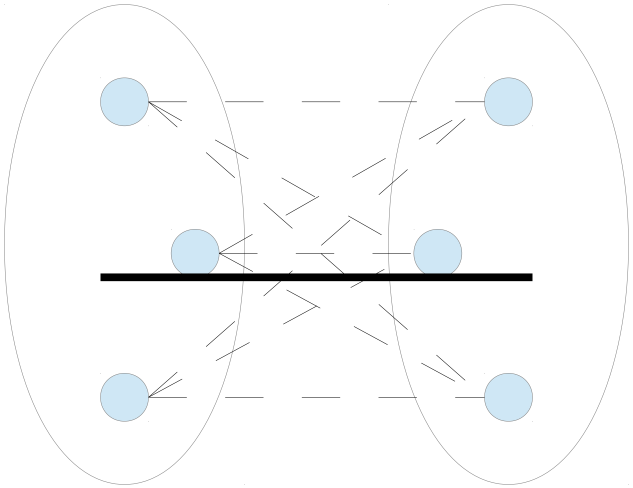
Hierarchical Agglomerative Clustering (HAC)

- Complete-link / complete-linkage clustering
  - Based on the similarity of the most dissimilar members

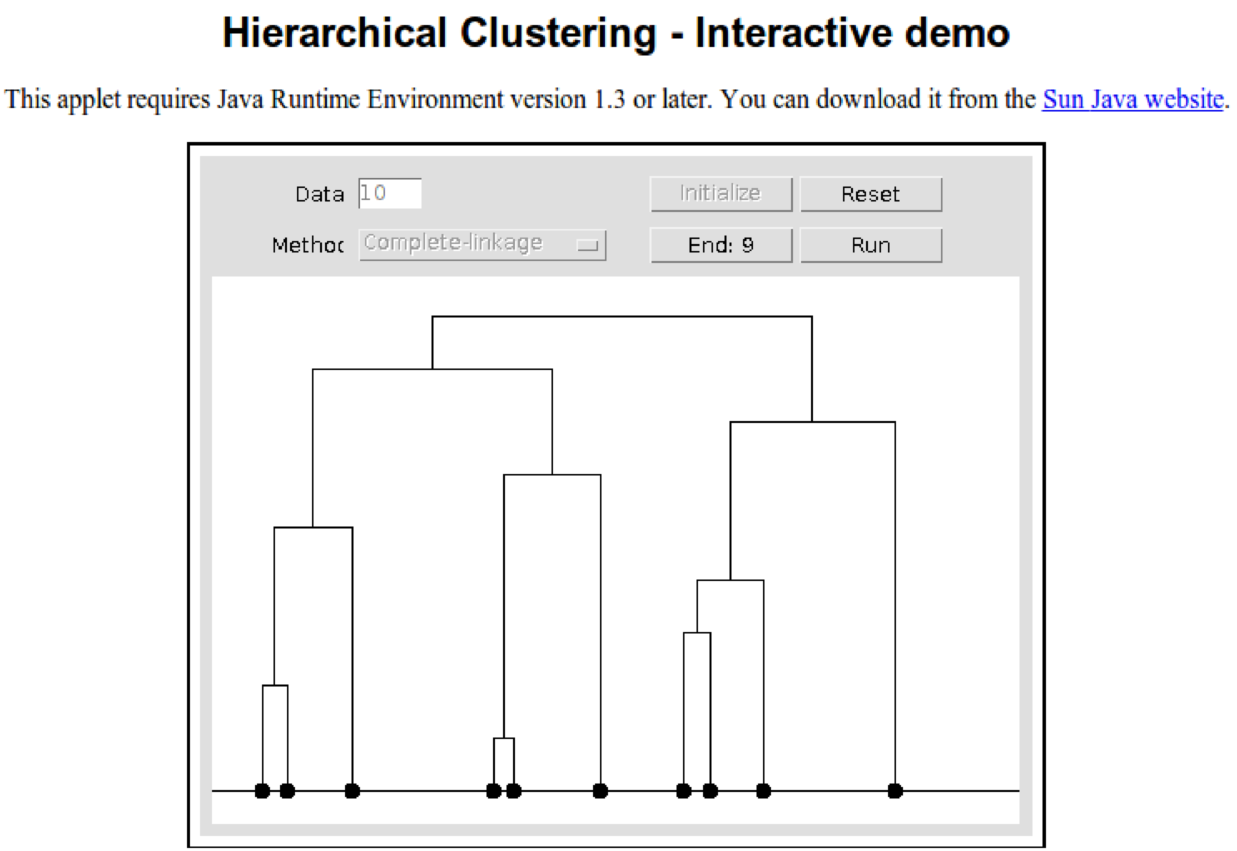


Hierarchical Agglomerative Clustering (HAC)

- Average-link / average-linkage clustering
  - Based on the average of all similarities between the members

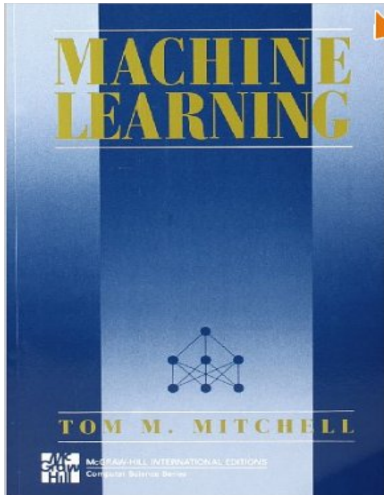


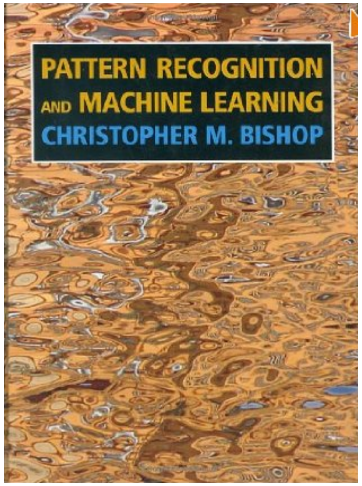
Hierarchical Agglomerative Clustering (HAC)

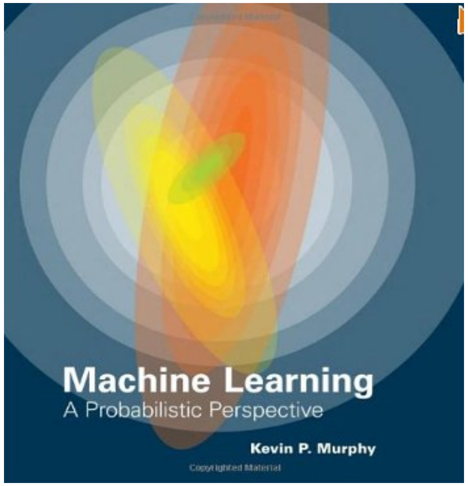


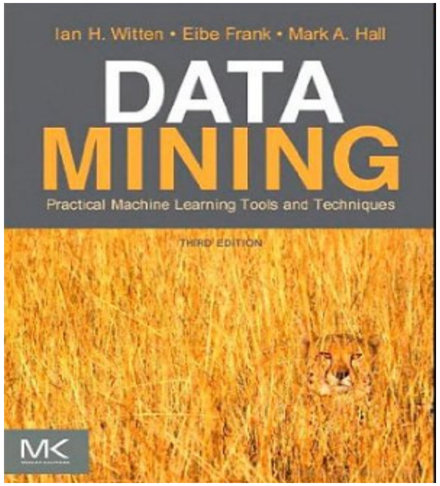
http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Further reading

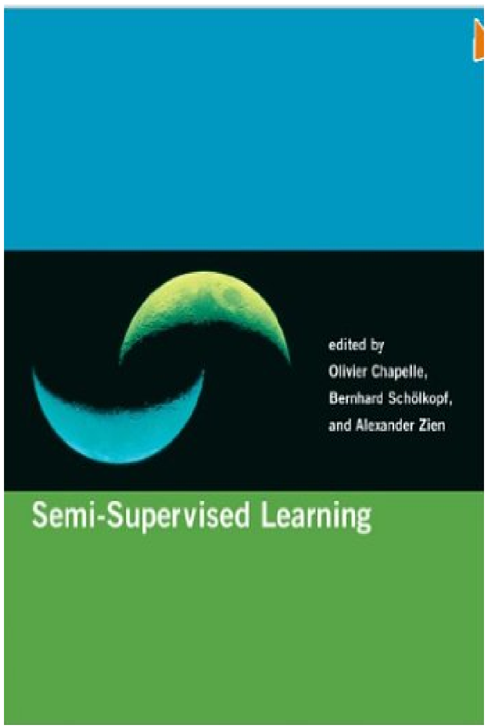


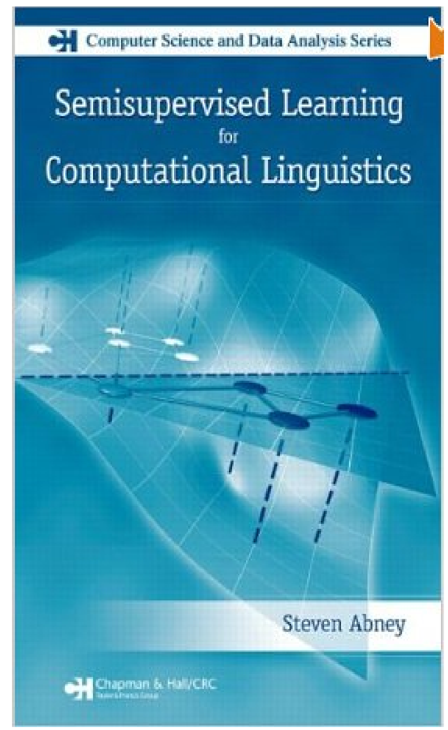






Further reading





Further reading