Statistical Natural Language Processing

Lecture 6: Machine Learning

Dr. Momtazi
Amirkabir University of Technology

Introduction

- Machine Learning
  - Field of study that gives computers the ability to learn without being explicitly programmed.

    [Arthur Samuel, 1959]

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

Outline

1. Supervised Learning

2. Semi-Supervised Learning

3. Unsupervised Learning

Outline

1. Supervised Learning

2. Semi-Supervised Learning

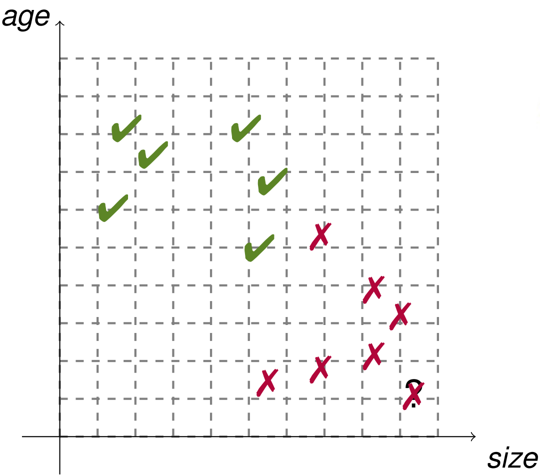
3. Unsupervised Learning

Supervised Learning

Renting budget: 1000 €

Size: 180 m^2
Age: 2 years ✗

Supervised Learning



Classification

Training

$T_1 \rightarrow C_1$
$T_2 \rightarrow C_2$
$\ldots$
$T_n \rightarrow C_n$

$\longrightarrow$

F_1
F_2
$\ldots$
F_n

$\longrightarrow$

$\boxed{\text{Model(F,C)}}$

Testing

$T_{n+1} \rightarrow ?$ $\longrightarrow$ F_{n+1} $\longrightarrow$ C_{n+1}

Applications

Problem	**Item**	**Category**
POS Tagging	Word	POS
Named Entity Recognition	Word	Named entity
Word Sense Disambiguation	Word	The 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

"I saw the man on the roof."

" $I_{[PRON]}$ $saw_{[V]}$ $the_{[DET]}$ $man_{[N]}$ $on_{[PREP]}$ $the_{[DET]}$ $roof_{[N]}$. "

[PRON]	Pronoun
[PREP]	Preposition
[DET]	Determiner
[V]	Verb
[N]	Noun
...	

Named Entity Recognition

"Steven Paul Jobs, co-founder of Apple Inc, was born in California."

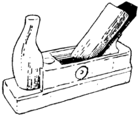
" $\underbrace{\textit{Steven Paul Jobs}}_{\textit{Person}}$, co-founder of $\underbrace{\textit{Apple Inc}}_{\textit{Organization}}$, was born in $\underbrace{\textit{California}}_{\textit{Location}}$."

Person
Organization
Location
Date
...

Word Sense Disambiguation

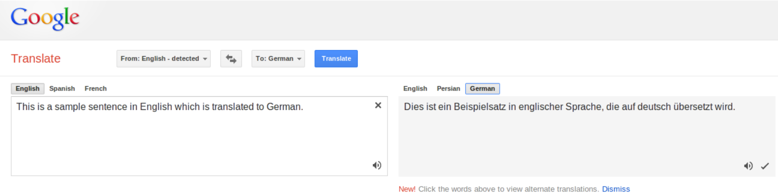
“Jim flew his plane to Texas.”

“Alice destroys the item with a plane.”

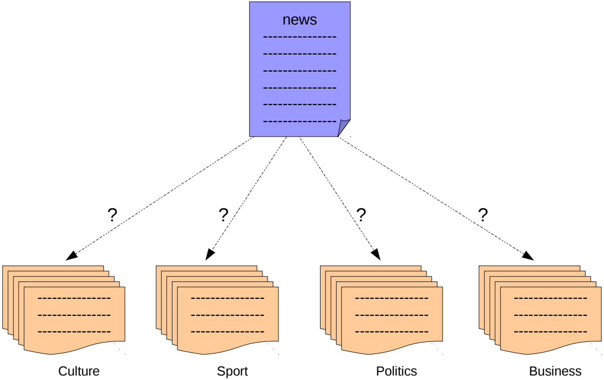


Spam Mail Detection

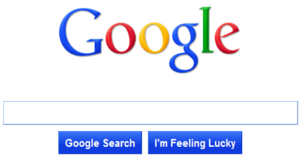
Language Identification



Text Categorization



Information Retrieval



Information technology - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Information_technology
Information Technology (IT) is concerned with technology to treat information. The acquisition, processing, storage and dissemination of vocal, pictorial, textual **...**
↳ Information systems - Information history - Category:Information technology

Information Technology – All About Information Technology - Wh...
jobsearchtech.about.com/od/careersintechnology/p/ITDefinition.htm
Information Technology and IT definition. What **information technology** actually means. How **information technology** is different from computer science.

RIT Information Sciences & Technology
www.ist.rit.edu/
offers bachelors and masters degrees in **information technology**, a masters degree in software development and management, and an advanced certificate in **...**

ScienceDaily: Information Technology News
www.sciencedaily.com/news/computers.../information_technology/
1 day ago – **Information Technology**. Read the latest in IT research from research institutes around the world. Updated daily, full-text, images, free.

Government of India, Department of Information Technology (DIT ...
www.mit.gov.in/
Developing the **information technology** industry. Includes an organisation chart, subsidiary bodies.

Information Technology - Everything You Need to Know
informationtechnology.net/
What is **Information Technology**? **Information Technology**, or IT, is the study, design, creation, utilization, support, and management of computer-based **...**

Information Technology
www.ibef.org/industry/informationtechnology.aspx
The Indian **information technology** (IT) industry has played a key role in putting India on the global map and is now envisioned to become a US$ 225 billion **...**

Information Technology - WetFeet.com
www.wetfeet.com/careers-industries/careers/information-technology
Information Technology. Overview. E-mail, personal computers, and the Internet: These products of the information age have become common currency among **...**

Classification

Training

$$\boxed{\begin{array}{l} T_1 \\ T_2 \\ \dots \\ T_n \end{array}} \begin{array}{l} \rightarrow C_1 \\ \rightarrow C_2 \\ \\ \rightarrow C_n \end{array} \longrightarrow \boxed{\begin{array}{c} F_1 \\ F_2 \\ \dots \\ F_n \end{array}} \longrightarrow \boxed{\text{Model(F,C)}}$$

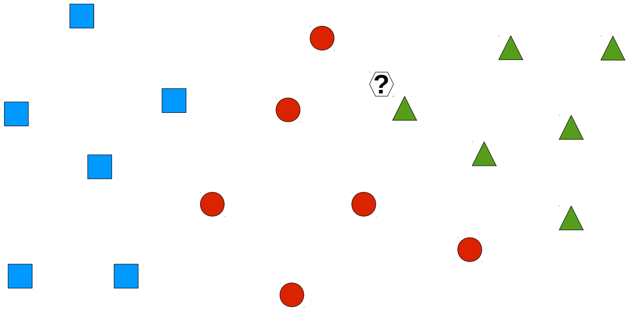
Testing

$$T_{n+1} \rightarrow ? \longrightarrow F_{n+1} \longrightarrow C_{n+1}$$

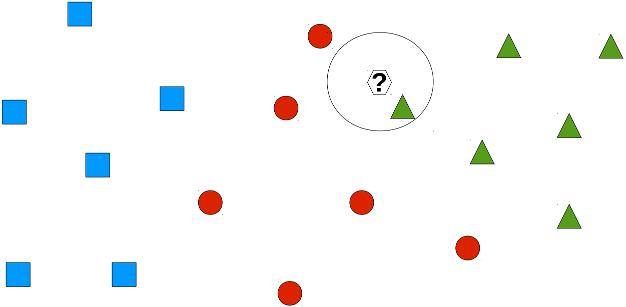
Classification Algorithms

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

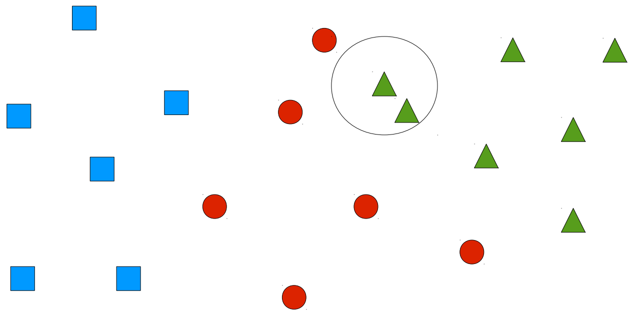
K Nearest Neighbor



K Nearest Neighbor

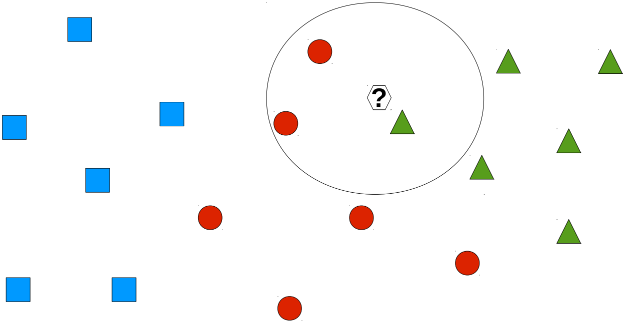


K Nearest Neighbor



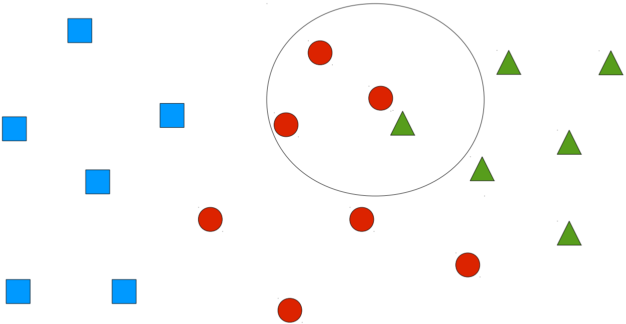
- 1 Nearest Neighbor

K Nearest Neighbor



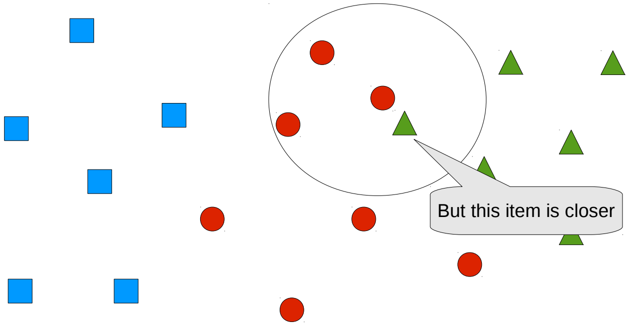
- 3 Nearest Neighbor

K Nearest Neighbor



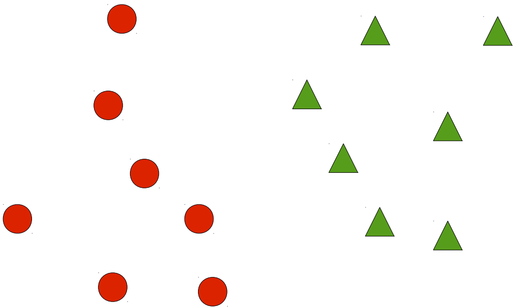
- 3 Nearest Neighbor

K Nearest Neighbor

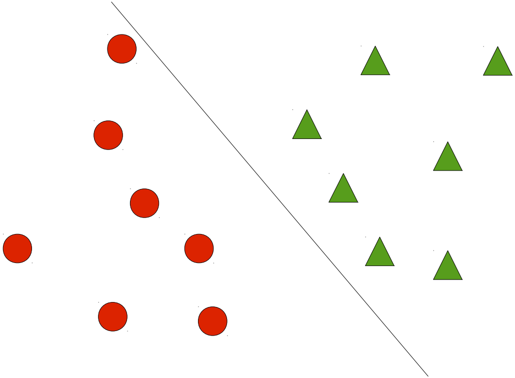


- 3 Nearest Neighbor

Support Vector Machines

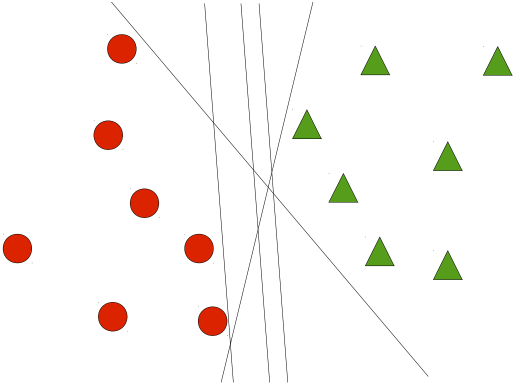


Support Vector Machines



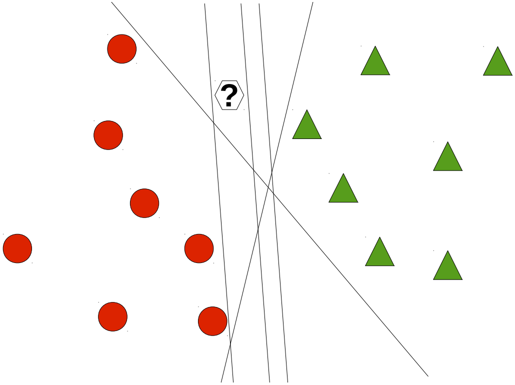
- Find a hyperplane in the vector space that separates the items of the two categories

Support Vector Machines



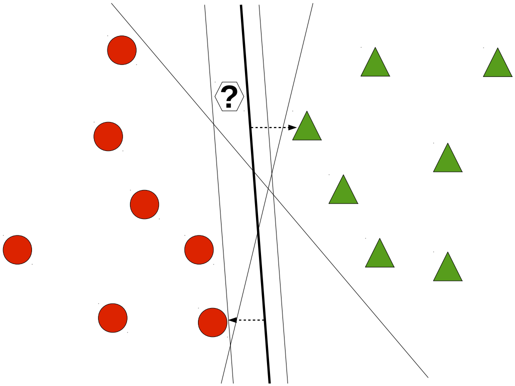
- There might be more than one possible separating hyperplane

Support Vector Machines



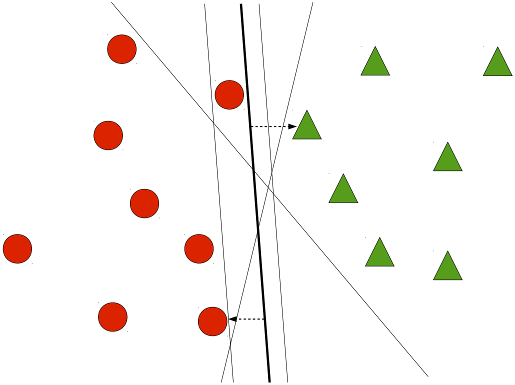
- There might be more than one possible separating hyperplane

Support Vector Machines



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

Support Vector Machines



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

Naïve Bayes

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

$$\hat{c} = \text{argmax}_{c_i} P(c_i)$$

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

$$P(c_i|d)$$

Naïve Bayes

$$\hat{c} = \mathrm{argmax}_{c_i} P(c_i)$$

$$\hat{c} = \mathrm{argmax}_{c_i} P(c_i|d)$$

$$\hat{c} = \mathrm{argmax}_{c_i} \frac{P(d|c_i) \cdot P(c_i)}{P(d)}$$

$P(d)$ has no effect

$$\hat{c} = \mathrm{argmax}_{c_i} P(d|c_i) \cdot P(c_i)$$

Naïve Bayes

$$\hat{c} = \text{argmax}_{c_i} P(d|c_i) \cdot P(c_i)$$

Likelihood Probability

Prior Probability

Maximum Entropy

- Assigning a weight λ_j to each feature f_j
  - Positive weight: the feature is likely to be effective
  - Negative weight: the feature is likely to be ineffective
- Picking out a subset of data by each feature
- Voting for each class based on the sum of weighted features

$$\hat{c} = \text{argmax}_{c_i} P(c_i|d, \lambda)$$

Maximum Entropy

$$\hat{c} = \text{argmax}_{c_i} P(c_i|d, \lambda)$$

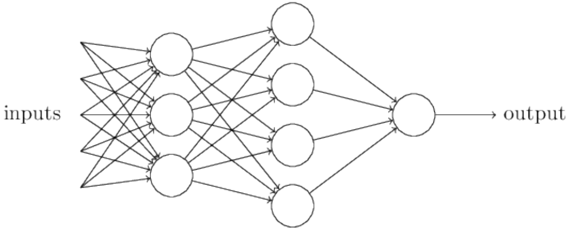
$$P(c_i|d, \lambda) = \frac{exp \sum_j \lambda_j \cdot f_j(c, d)}{\sum_{c_i} exp \sum_j \lambda_j \cdot f_j(c_i, d)}$$

- The expectation of each feature is calculated as follows:

$$E(f_i) = \sum_{(c,d) \in (C,D)} P(c, d) \cdot f_i(c, d)$$

Neural Networks

- Mainly inspired by human brain
- A Human brain is
  - Capable of computationally demanding perceptual acts - like, recognition of faces, speech
  - Highly parallel computing structure
  - Imprecise information processing
  - Collection of more than 10 billion interconnected neuron

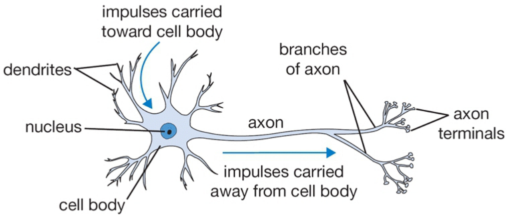


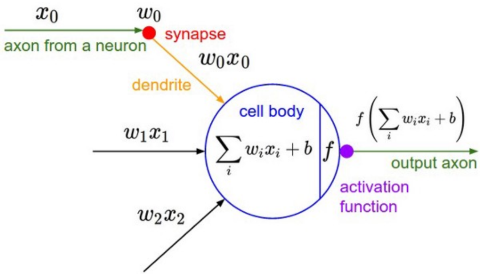
Neural Networks

- Neuron: the basic unit of computation in a neural network (also called a node or unit)

- Architecture
  - Input Nodes (input layer)
  - Hidden nodes (hidden layer)
  - Output Nodes (output layer)
  - Connections and weights
  - Activation function
  - Learning rule

Neural Networks





Neural Networks

- Input Nodes (input layer):
  - No computation is done here within this layer
  - They just pass the information to the next layer.
  - A block of nodes is also called layer.

Neural Networks

- Hidden nodes (hidden layer):
  - Hidden layers is where intermediate processing or computation is done.
  - They perform computations and then transfer the weights (signals or information) from the input layer to the following layer.
  - Next layer is another hidden layer or to the output layer.

Neural Networks

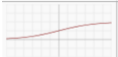
- Output Nodes (output layer):
  - Here we finally use an activation function that maps to the desired output format.

Neural Networks

- Connections and weights:
  - The network consists of connections, each connection transferring the output of a neuron to the input of another neuron.
  - Each connection is assigned a weight assigned on the basis of its relative importance to other inputs.

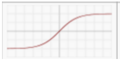
Neural Networks

- Activation function:
  - The activation function of a node defines the output of that node given an input or set of inputs (also called the transfer function).
  - It introduces non-linearity into the output of a neuron.
  - Taking a single number and performing a certain fixed mathematical operation on it.
  - Types:
    - Sigmoid



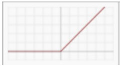
$$f(x) = \frac{1}{1+e^{-x}}$$

    - Tanh



$$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

    - ReLU



$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

    - Softmax

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

Neural Networks

- Learning rule:
  - The learning rule is a rule or an algorithm which modifies the parameters of the neural network, in order to produce a favored output given an input to the network.
  - The learning process typically amounts to modifying the weights and thresholds.

Classification

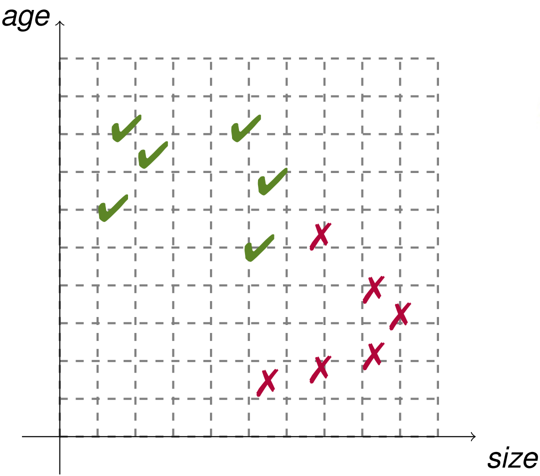
Training

$$
\begin{array}{l} T_1 \rightarrow C_1 \\ T_2 \rightarrow C_2 \\ \dots \\ T_n \rightarrow C_n \end{array}
\longrightarrow
\begin{array}{c} F_1 \\ F_2 \\ \dots \\ F_n \end{array}
\longrightarrow
\boxed{\text{Model(F,C)}}
$$

Testing

$$
T_{n+1} \rightarrow ? \longrightarrow F_{n+1} \longrightarrow C_{n+1}
$$

Feature Selection



Parking
Number of rooms
Balcony
...

Feature Selection

- Bag-of-words:
  - Each document can be represented by the set of words that appear in the document
  - Result is a high dimensional feature space
  - The process is computationally expensive

- Solution
  - Using a feature selection method to select informative words

Feature Selection Methods

- Information Gain
- Mutual Information
- χ-Square

Information Gain

- An entropy-based feature evaluation method
- Measuring the number of bits required for category prediction w.r.t. the presence or absence of a term in the document
- Removing words whose information gain is less than a predefined threshold

$$
\begin{aligned}
IG(w) = & -\sum_{i=1}^{K} P(c_i) \log P(c_i) \\
& + P(w) \sum_{i=1}^{K} P(c_i|w) \log P(c_i|w) \\
& + P(\overline{w}) \sum_{i=1}^{K} P(c_i|\overline{w}) \log P(c_i|\overline{w})
\end{aligned}
$$

Information Gain

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

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

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

N: # docs
N_i: # docs in category c_i
N_w: # docs containing w
$N_{\overline{w}}$: # docs not containing w
N_{iw}: # docs in category c_i containing w
$N_{i\overline{w}}$: # docs in category c_i not containing w

Mutual Information

- Measuring the effect of each word in predicting the category
  - How much does its presence or absence in a document contribute to category prediction?

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

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

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

$$MI(w) = \sum_i P(c_i) \cdot MI(w, c_i)$$

χ-square

- Measuring the dependencies between words and categories

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

- 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
  - POS Tagging
  - Named Entity Recognition

(will be discussed later)

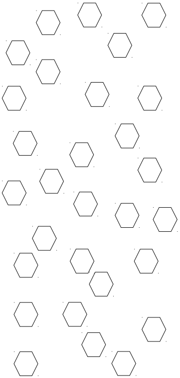
Shortcoming

- Data annotation is labor intensive

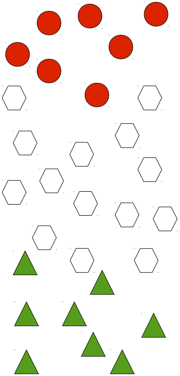
- Solution:
  - Using a minimum amount of annotated data
  - Annotating further data by human, if they are very informative

Active Learning

Active Learning

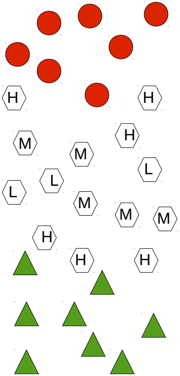


Active Learning



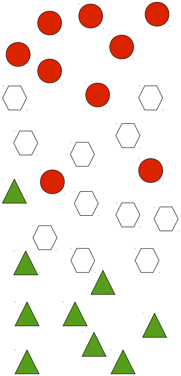
- Annotating a small amount of data

Active Learning



- Annotating a small amount of data
- Calculating the confidence score of the classifier on unlabeled data

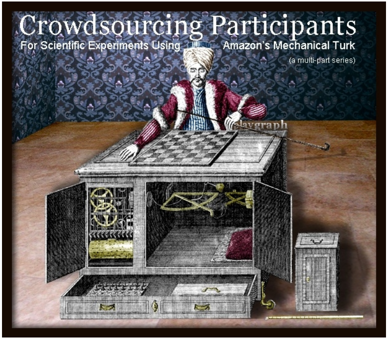
Active Learning



- Annotating a small amount of data
- Calculating the confidence score of the classifier on unlabeled data
- Finding the informative unlabeled data (data with lowest confidence)
- Annotating the informative data by the human

Active Learning

Amazon Mechanical Turk



Outline

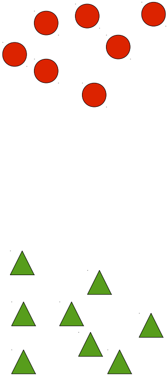
1. Supervised Learning
2. Semi-Supervised Learning
3. Unsupervised Learning

Semi-Supervised Learning

- Problem of data annotation

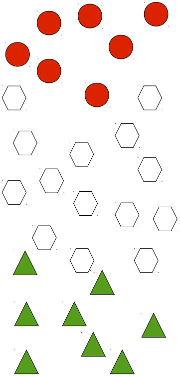
- Solution:
  - Using minimum amount of annotated data
  - Annotating further data **automatically**, if they are easy to predict

Semi-Supervised Learning



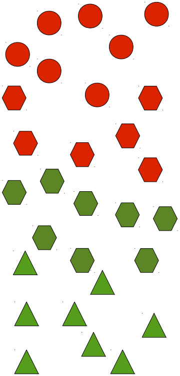
- A small amount of labeled data

Semi-Supervised Learning



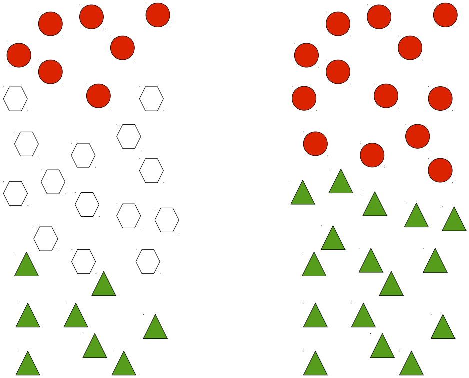
- A small amount of labeled data
- A large amount of unlabeled data

Semi-Supervised Learning



- A small amount of labeled data
- A large amount of unlabeled data
- Solution
  - Finding the similarity between the labeled and unlabeled data
  - Predicting the labels of the unlabeled data

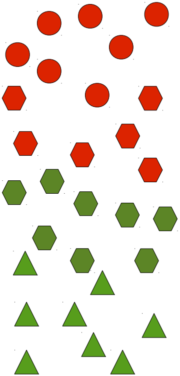
Semi-Supervised Learning



- Training the classifier using
  - The labeled data
  - Predicted labels of the unlabeled data

Shortcoming

- Introducing a lot of noisy data to the system

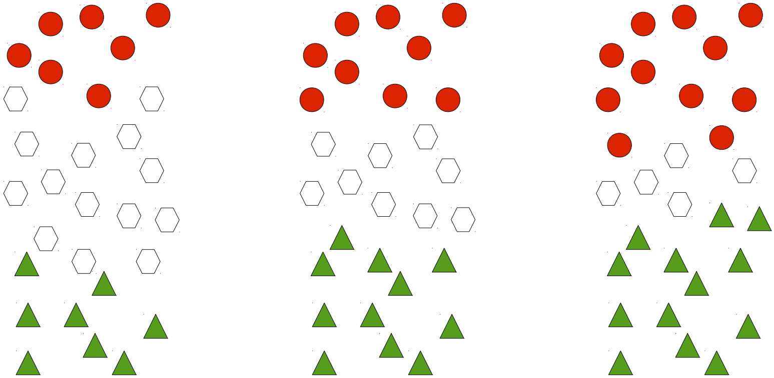


Shortcoming

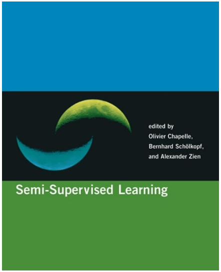
- Introducing a lot of noisy data to the system

- Solution
  - Adding unlabeled data to the training set, if the predicted label has a high confidence

Semi-Supervised Learning



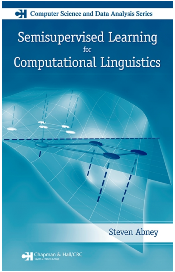
Related Books



Semi-Supervised Learning

by O. Chapelle, B. Schölkopf, A. Zien
MIT Press

2006



Semisupervised Learning for Computational Linguistics

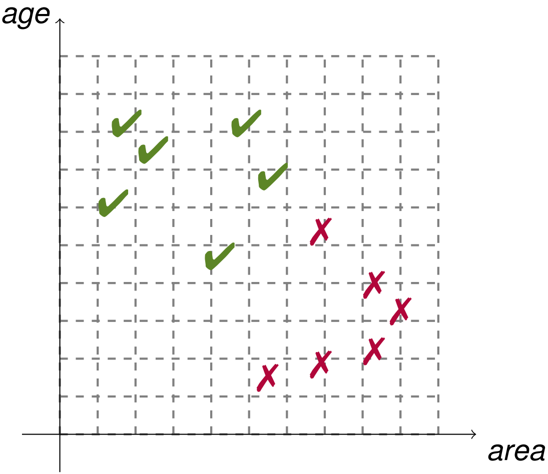
by S. Abney
Chapman & Hall

2007

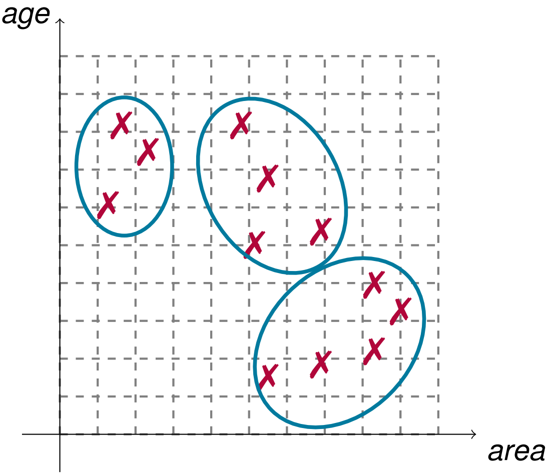
Outline

1. Supervised Learning
2. Semi-Supervised Learning
3. Unsupervised Learning

Unsupervised Learning



Unsupervised Learning



Clustering

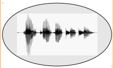
- Working based on the similarities between the data items
- Assigning the similar data items to the same cluster

Applications

- Word Clustering
  - Speech Recognition
  - Machine Translation
  - Named Entity Recognition
  - Information Retrieval
  - ...

- Document Clustering
  - Information Retrieval
  - Social Network Analysis
  - ...

Speech recognition



⇒ *"Computers can recognize speech."*

"Computers can wreck a nice peach."

Machine Translation

"The cat eats ..."	$\Rightarrow$	*"Die Katze frisst ..."*
		"Die Katze isst ..."

Language Modeling

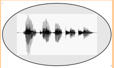
Corpus Texts:

"I have a meeting on Monday evening"

"You should work on Wednesday afternoon"

"The next session of the NLP lecture in on Thursday morning"

no observation in the corpus.



⇒ *"The talk is on Monday morning."*

"The talk is on Monday molding."

Saturday
Wednesday
Monday
Thursday
Tuesday
Friday
Sunday

afternoon
morning
evening
night

Language Modeling

Corpus Texts:

"I have a meeting on Monday evening"

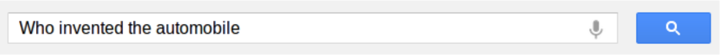
"You should work on Wednesday afternoon"

"The next session of the NLP lecture in on Thursday morning"

⇒ [Week-day] [day-time]

Class-based Language Model

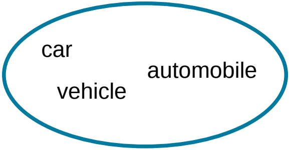
Information Retrieval



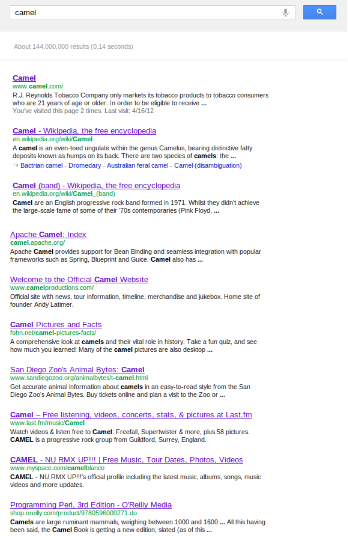
"The first car was invented by Karl Benz."

"Thomas Edison invented the first commercially practical light."

"Alexander Graham Bell invented the first practical telephone."



Information Retrieval



camel

About 144,000,000 results (0.14 seconds)

Camel
www.**camel**.com/
R.J. Reynolds Tobacco Company only markets its tobacco products to tobacco consumers who are 21 years of age or older. In order to be eligible to receive ...
You've visited this page 2 times. Last visit: 4/16/12

Camel - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Camel**
A **camel** is an even-toed ungulate within the genus Camelus, bearing distinctive fatty deposits known as humps on its back. There are two species of **camels**: the ...
↳ Bactrian camel - Dromedary - Australian feral camel - Camel (disambiguation)

Camel (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Camel**_(band)
Camel are an English progressive rock band formed in 1971. Whilst they didn't achieve the large-scale fame of some of their '70s contemporaries (Pink Floyd, ...

Apache **Camel**: Index
camel.apache.org/
Apache **Camel** provides support for Bean Binding and seamless integration with popular frameworks such as Spring, Blueprint and Guice. **Camel** also has ...

Welcome to the Official **Camel** Website
www.**camel**productions.com/
Official site with news, tour information, timeline, merchandise and jukebox. Home site of founder Andy Latimer.

Camel Pictures and Facts
fohn.net/**camel**-pictures-facts/
A comprehensive look at **camels** and their vital role in history. Take a fun quiz, and see how much you learned! Many of the **camel** pictures are also desktop ...

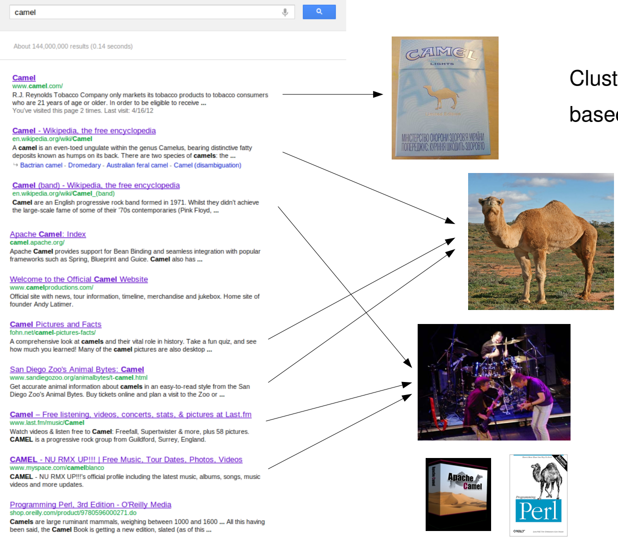
San Diego Zoo's Animal Bytes: **Camel**
www.sandiegozoo.org/animalbytes/t-**camel**.html
Get accurate animal information about **camels** in an easy-to-read style from the San Diego Zoo's Animal Bytes. Buy tickets online and plan a visit to the Zoo or ...

Camel – Free listening, videos, concerts, stats, & pictures at Last.fm
www.last.fm/music/**Camel**
Watch videos & listen free to **Camel**: Freefall, Supertwister & more, plus 58 pictures. **CAMEL** is a progressive rock group from Guildford, Surrey, England.

CAMEL - NU RMX UP!!! | Free Music, Tour Dates, Photos, Videos
www.myspace.com/**camel**blanco
CAMEL - NU RMX UP!!!'s official profile including the latest music, albums, songs, music videos and more updates.

Programming Perl, 3rd Edition - O'Reilly Media
shop.oreilly.com/product/9780596000271.do
Camels are large ruminant mammals, weighing between 1000 and 1600 ... All this having been said, the **Camel** Book is getting a new edition, slated (as of this ...

Information Retrieval


camel

About 144,000,000 results (0.14 seconds)

Camel
www.camel.com/
R.J. Reynolds Tobacco Company only markets its tobacco products to tobacco consumers who are 21 years of age or older. In order to be eligible to receive ...
You've visited this page 2 times. Last visit: 4/16/12

Camel - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Camel
A **camel** is an even-toed ungulate within the genus Camelus, bearing distinctive fatty deposits known as humps on its back. There are two species of **camels**: the ...
Bactrian camel - Dromedary - Australian feral camel - Camel (disambiguation)

Camel (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Camel_(band)
Camel are an English progressive rock band formed in 1971. Whilst they didn't achieve the large-scale fame of some of their '70s contemporaries (Pink Floyd, ...

Apache Camel: Index
camel.apache.org/
Apache **Camel** provides support for Bean Binding and seamless integration with popular frameworks such as Spring, Blueprint and Guice. **Camel** also has ...

Welcome to the Official Camel Website
www.camelproductions.com/
Official site with news, tour information, timeline, merchandise and jukebox. Home site of founder Andy Latimer.

Camel Pictures and Facts
fohn.net/camel-pictures-facts/
A comprehensive look at **camels** and their vital role in history. Take a fun quiz, and see how much you learned! Many of the **camel** pictures are also desktop ...

San Diego Zoo's Animal Bytes: Camel
www.sandiegozoo.org/animalbytes/t-camel.html
Get accurate animal information about **camels** in an easy-to-read style from the San Diego Zoo's Animal Bytes. Buy tickets online and plan a visit to the Zoo or ...

Camel – Free listening, videos, concerts, stats, & pictures at Last.fm
www.last.fm/music/Camel
Watch videos & listen free to **Camel**: Freefall, Supertwister & more, plus 58 pictures. **CAMEL** is a progressive rock group from Guildford, Surrey, England.

CAMEL - NU RMX UP!!! | Free Music, Tour Dates, Photos, Videos
www.myspace.com/camelblanco
CAMEL - NU RMX UP!!!'s official profile including the latest music, albums, songs, music videos and more updates.

Programming Perl, 3rd Edition - O'Reilly Media
shop.oreilly.com/product/9780596000271.do
Camels are large ruminant mammals, weighing between 1000 and 1600 ... All this having been said, the **Camel** Book is getting a new edition, slated (as of this ...

Clustering documents

based on their similarities

Tweet Clustering

video : **bitcoin : mtgox** exchange goes offline - **bitcoin** , a virtual currency ... the slow-motion collapse of **mt . gox** is **bitcoin**'s first financial crisis ... Disastro **bitcoin** : **mt . gox** cessa ogni attivite ... **: mt . gox** , il pi ù grande cambia
california couple finds time capsules **worth $10 million** **californian couple finds $10 million worth** of gold coins in tin can
ukraine puts off vote on new government despite eu pleas for quick action - washing **ukraine** truce shattered , death toll hits 67 - kiev (reuters) - ukraine suffered its bloo **ukraine** fighting leaves at least 18 dead as kiev barricades burn - clashes in ukraine ..
are you going to come on his network and get poor ratings too **?** **are you** sold on the waffle taco **?**
the chromecast app flood has started **by** **the** importance of emotion in design **by**

Calculating Words' Similarities

- Representing words as vector and using vector similarity measures (e.g., cosine similarity)
  - Word/Doc matrix: frequency of words in documents
    - LSA
    - LDA
  - Word context: co-occurrence of words with other words in a window of text
    - Word2Vec
    - Glove

- Using word semantic similarity from available semantic resources (e.g., WordNet)

Calculating Documents' Similarities

- Representing documents as vector and using vector similarity measures (e.g., cosine similarity)
  - Word/Doc matrix: frequency of words in documents; TF-IDF
    - LSA
    - LDA
  - More contextual information
    - Doc2Vec

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
- Works well for many cases
- Used as default / baseline for clustering documents

- Algorithm
  - Defining each cluster center as the mean or centroid of the items in the cluster

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

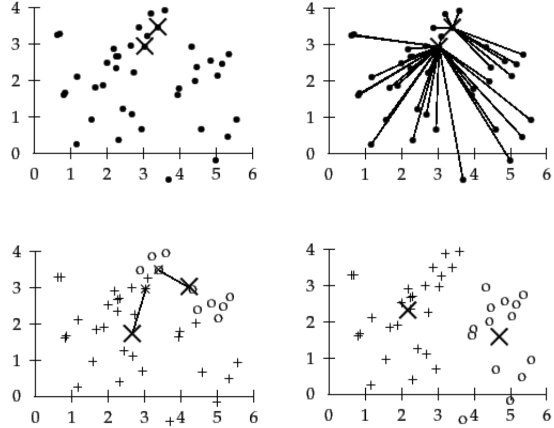
  - Minimizing the average squared Euclidean distance of the items from their cluster centers

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
```

- 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

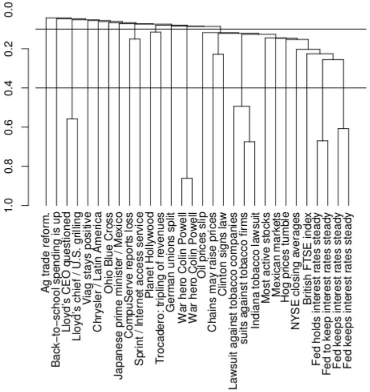
K-means



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

Hierarchical Agglomerative Clustering (HAC)

- Creating a hierarchy in the form of a binary tree



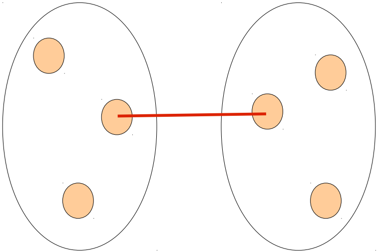
Hierarchical Agglomerative Clustering (HAC)

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

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

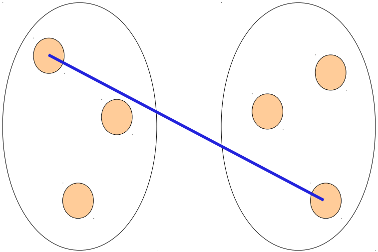
Hierarchical Agglomerative Clustering (HAC)

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



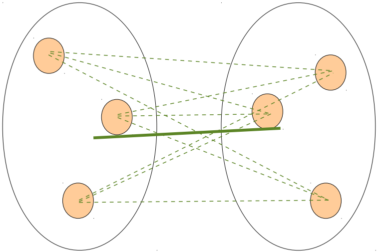
Hierarchical Agglomerative Clustering (HAC)

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

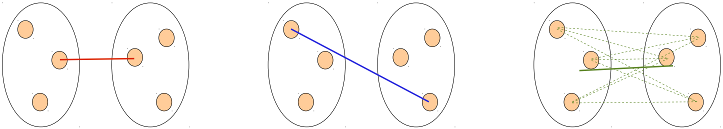


Hierarchical Agglomerative Clustering (HAC)

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



Hierarchical Agglomerative Clustering (HAC)



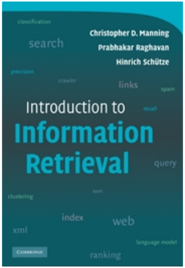
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Further Reading

Introduction to Information Retrieval

C.D. Manning, P. Raghavan, H. Schütze Cambridge

University Press 2008



`http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html`

Chapters 13,14,15,16,17