ELEG 6913: Machine Learning for Big Data

Fall 2016

Lecture 8: Feature Engineering on Texts

Dr. Xishuang Dong

Outline

- **Text Representation**
- **Feature Extraction**
- **Feature Selection**
- **Summary**

(Acknowledgment: some parts of the slides are from Guoping Qiu, Jen Golbeck, and various other sources. The copyright of those parts belongs to their original owners.)

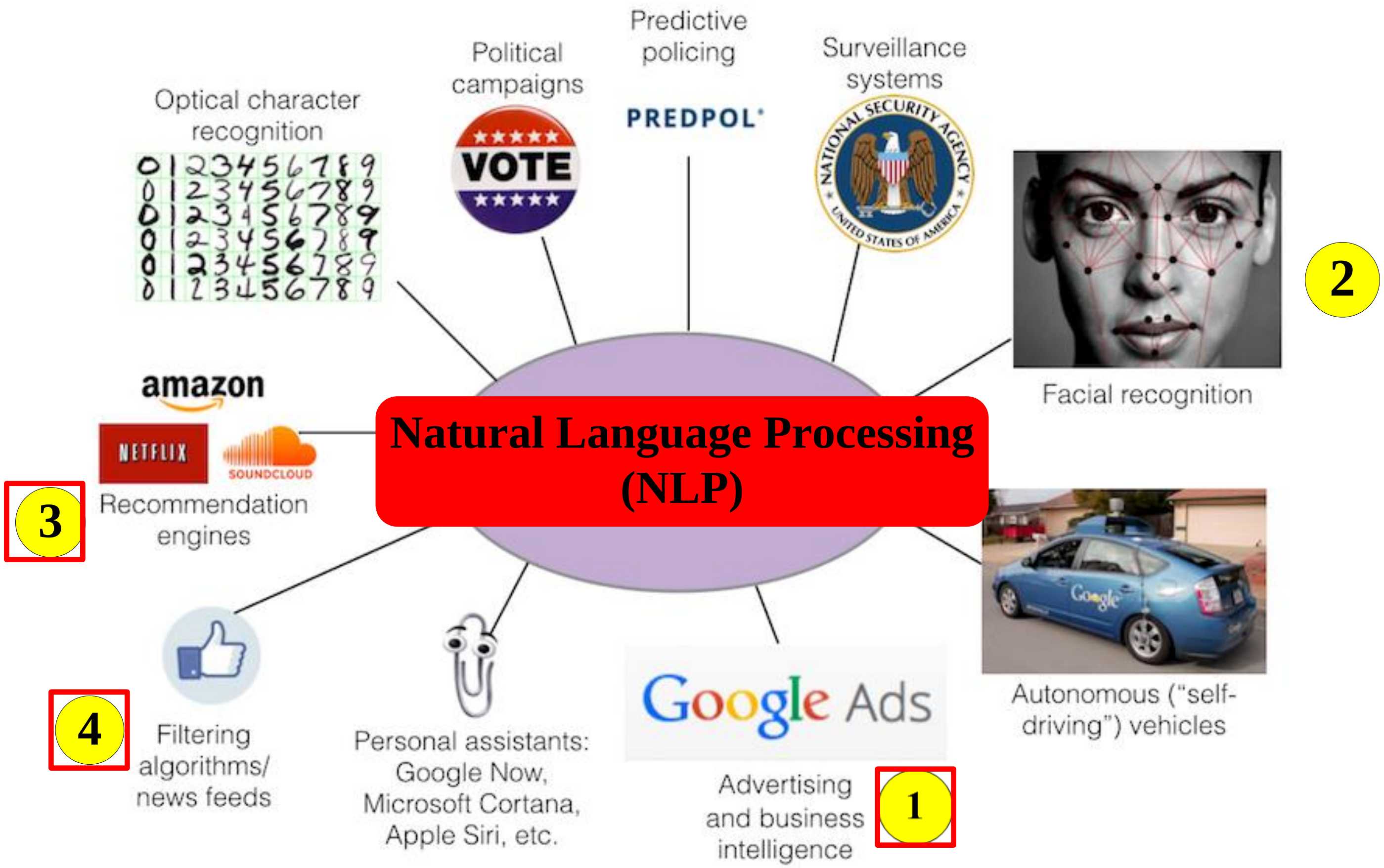
Outline

- **Text Representation**
- **Feature Extraction**
- **Feature Selection**
- **Summary**

Machine Learning

- **Machine Learning and Applications**
- **Data Representation**
- **Text Representation**

Machine Learning Applications



https://redshiftzero.github.io/2015/08/29/Manipulation-and-Machine-Learning/

Machine Learning Problems

- **Classification**
- **Clustering**
- **Sequence Forecasting**
- **… …**

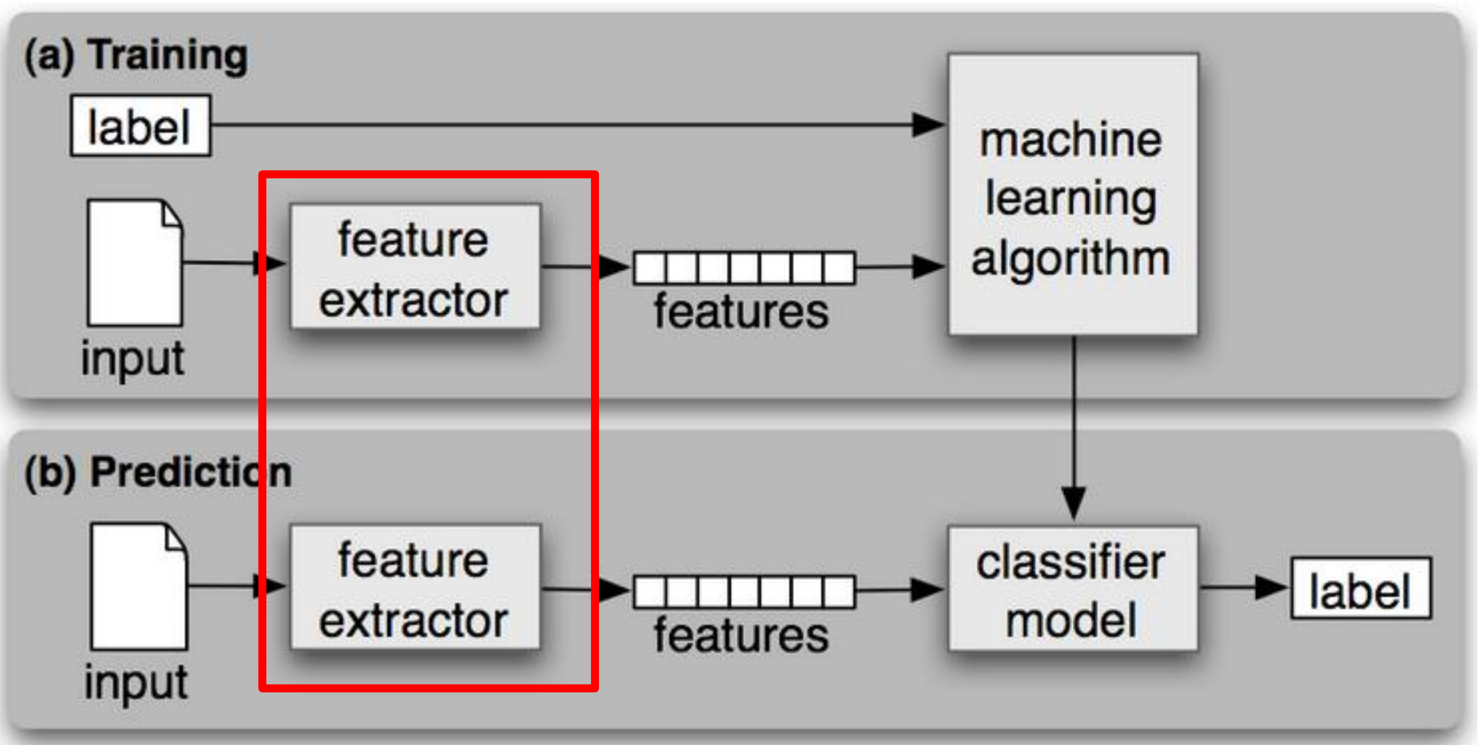
Natural Language Processing

- **Text Classification**
- **Text Clustering**

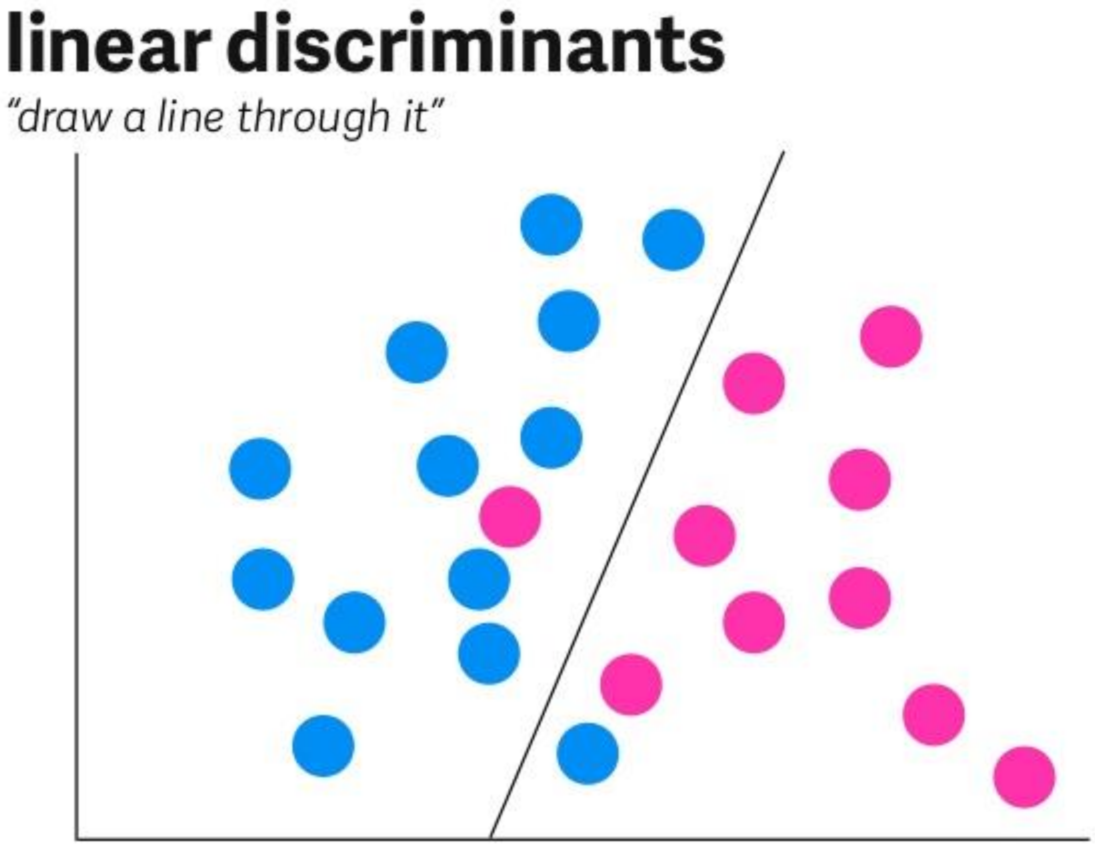
Applications of Text Classification

- Information Retrieval
- Question Answer (Q & A)
- Recommendation System
- Stock Prediction
- Suicide Forecasting
- … …

Supervised Machine Learning



Data Representation



Data Representation

Training sample pairs (X, D)

$X = (x_1, x_2, \ldots, x_n)$ is the feature vector representing the instance.

$D = (d_1, d_2, \ldots d_m)$ is the desired (target) output of the classifier

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

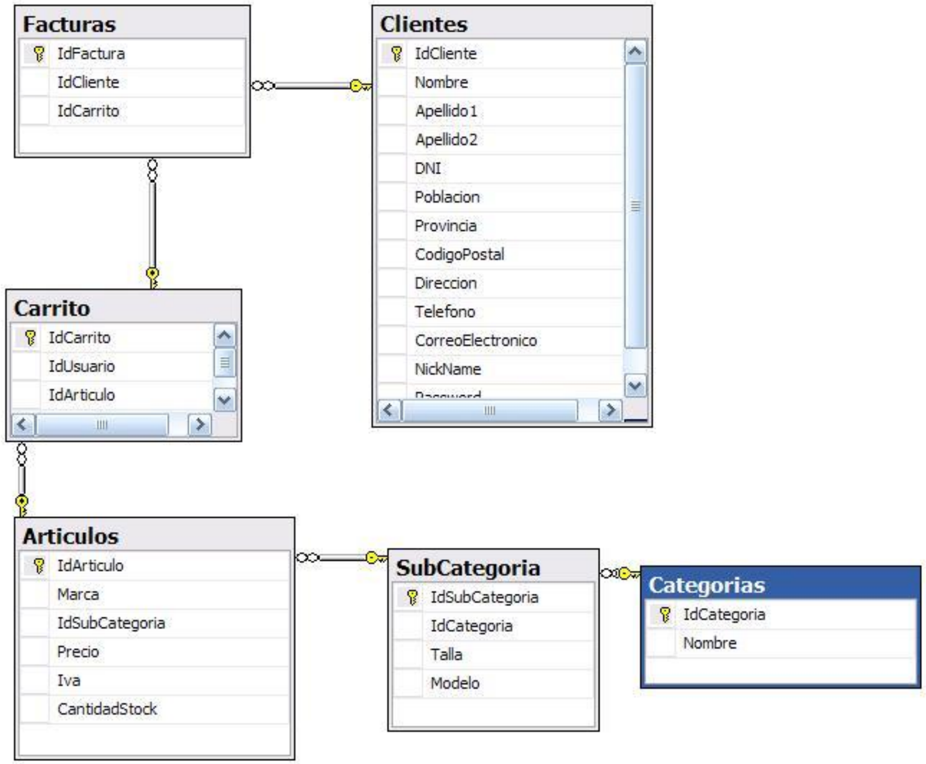
Data Representation

- **Structured Data**
- **Unstructured Data**

Structured Data

- **Data structure is a particular way of organizing data in a computer so that it can be used efficiently (Wikipedia).**
- **A logic model of a particular organization.**

Tables in Database



Tables in Database

Training sample pairs (X, D)

X = $(x_1, x_2, \ldots, x_n)$ is the feature vector representing the instance.

D= $(d_1, d_2, \ldots d_m)$ is the desired (target) output of the classifier

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Unstructured Data

- **Text**

  "Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome…"

Unstructured Data

How to represent texts for learning in computers?

Text Representation

- **Where can we gain the texts?**
- **How do you choose the source of the texts?**
- **Scale of the text datasets**

Where can we gain the texts?

- **Datasets Library**

About Citation Policy Donate a Data Set Contact

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Search

Repository Web Google

View ALL Data Sets

Browse Through: **30** Data Sets

Table View List View

Default Task

Classification (21)
Regression (4)
Clustering (9)
Other (5)

Attribute Type

Categorical (2)
Numerical (13)
Mixed (0)

Data Type - Undo

Multivariate (274)
Univariate (16)
Sequential (35)
Time-Series (63)
Text (30)
Domain-Theory (22)
Other (21)

Area

Life Sciences (2)
Physical Sciences (1)
CS / Engineering (14)
Social Sciences (2)
Business (3)

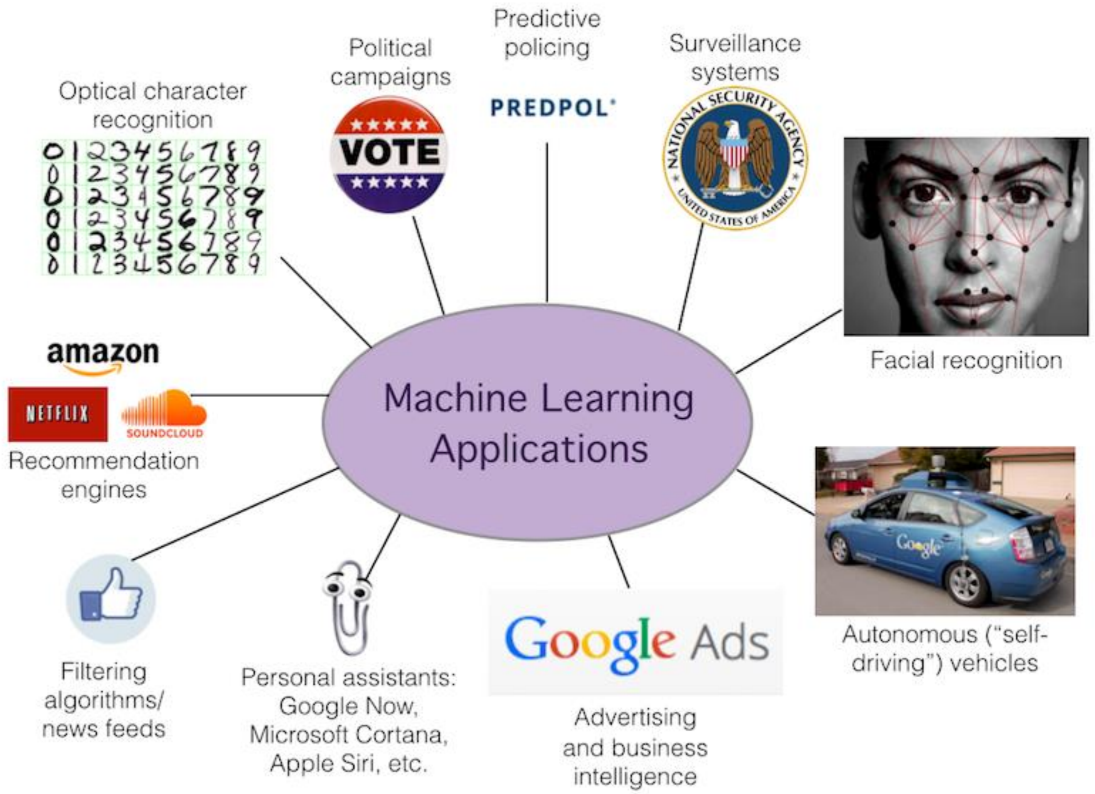
Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
3D Road Network (North Jutland, Denmark)	Sequential, Text	Regression, Clustering	Real	434874	4	2013
Amazon Commerce reviews set	Multivariate, Text, Domain-Theory	Classification	Real	1500	10000	2011
Badges	Univariate, Text	Classification		294	1	1994
Bag of Words	Text	Clustering	Integer	8000000	100000	2008
CNAE-9	Multivariate, Text	Classification	Integer	1080	857	2012
DBWorld e-mails	Text	Classification		64	4702	2011

Where can we gain the texts?

- **Collect from website**
  - ✓ **Blogs**
  - ✓ **E-commerce websites**
  - ✓ **Social Network**
    - ➢ **Twitter**
  - ✓ **….**

How do you choose the source of the texts?

- **Applications determine the source.**



Scale of the text datasets

- Infinite data is best, but…
- Standard Machine Learning Experiments
  - ✓ More than 1,000 samples
- Big Data
  - ✓ More than 1 Gigbyte
  - ✓ More than 1 million samples (Deep Learning)

Outline

- Text Representation
- **Feature Extraction**
- Feature Selection
- Summary

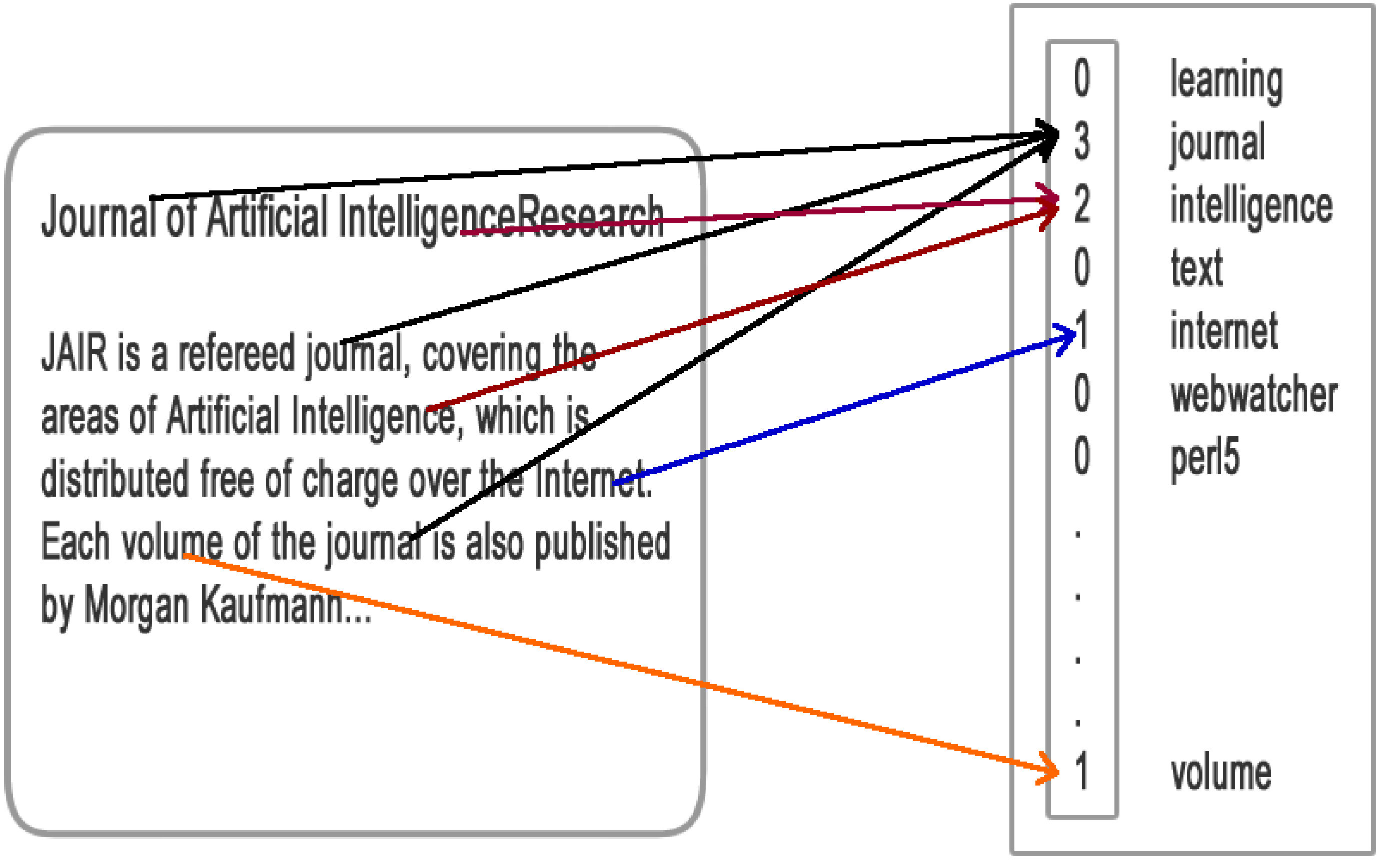
Feature Extraction

- **Bag of words**
- **Binary Features**
- **Continuous Features**

Bag of words

- **Vector representation doesn't consider the ordering of words in a document**
- **In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order.**

Bag of words


Journal of Artificial IntelligenceResearch
JAIR is a refereed journal, covering the areas of Artificial Intelligence, which is distributed free of charge over the Internet. Each volume of the journal is also published by Morgan Kaufmann...
0 learning
3 journal
2 intelligence
0 text
1 internet
0 webwatcher
0 perl5
1 volume

Binary Features

D1: John likes to watch movies. Mary likes movies too.

D2: John also likes to watch football games.

```
[
    "John",
    "likes",
    "watch",
    "movies",
    "also",
    "football",
    "games",
    "Mary",
    "too"
]
```

Binary Representation

D1: [1, 1, 1, 1, 1, 0, 0, 0, 1, 1]

D2: [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Bag of words

- Deficiency
  - ✓ Week representation
  - ✓ Sparse representation

Binary Features

D1: John doesn’t like to watch movies, but likes to watch football games.

D2: John doesn’t like to watch football games, but likes to watch movies.

Binary Features

D1: John doesn’t like to watch movies. Mary likes to watch football games.

D2: John likes singing.

Continuous Features

- **Term Frequency (TF)**
- **Term Frequency-Inverse Document Frequency (TF-IDF)**

Term Frequency (TF)

- **The term frequency $tf_{t,d}$ of term *t* in document *d* is defined as the number of times that *t* occurs in *d*.**
- **A document with 10 occurrences of the term is more relevant than a document with one occurrence of the term.**
- **But not 10 times more relevant.**
- **Relevance does not increase proportionally with term frequency**

Term Frequency (TF)

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

[

a , 3,

the, 2

or, 2

… …

]

Stop Words

Term Frequency

D1: John likes to watch movies. Mary likes movies too.

D2: John also likes to watch football games.

[
"John",
"likes",
"watch",
"movies",
"also",
"football",
"games",
"Mary",
"too"
]

Term Frequency Representation

D1: [1, 2, 1, 2, 1, 0, 0, 0, 1, 1]

D2: [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Term Frequency-Inverse Document Frequency (TF-IDF)

- **Document frequency**
  - ✓ **Rare terms are more informative than frequent terms.**
    - ➢ **Recall stop words**
  - ✓ **We want a high weight for rare terms like “arachnocentric”.**

Term Frequency-Inverse Document Frequency (TF-IDF)

- **Document frequency**
  - ✓ **A document containing such a term is more likely to be relevant than a document that doesn’t, but it’s not a sure indicator of relevance.**
  - ✓ **We will use document frequency (df) to capture this in the score.**

df ($\leq N$) is the number of documents that contain the term

Term Frequency-Inverse Document Frequency (TF-IDF)

- **Document frequency**
  - ✓ ***df_t* is the document frequency of *t*: the number of documents that contain *t***
    - ➢ ***df* is a measure of the informativeness of *t*.**

Term Frequency-Inverse Document Frequency (TF-IDF)

- **Document frequency**
  - ✓ **We define the IDF (inverse document frequency) of *t* by**

$$\mathrm{idf}_t = \log_{10} N/\mathrm{df}_t$$

N: the number of documents

- **TF-IDF**
  - ✓ ***TFIDF(t) = TF(t)*IDF(t)***

Outline

- Text Representation
- Feature Extraction
- **Feature Selection**
- Summary

Feature Selection

- **In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.**

Feature Selection

- **Feature selection techniques are used for three reasons:**
  - ✓ **simplification of models to make them easier to interpret by researchers/users**
  - ✓ **shorter training times,**
  - ✓ **enhanced generalization by reducing overfitting**

Information Gain

- **In information theory and machine learning, information gain is a synonym for Kullback–Leibler divergence that is a measure of the difference between two probability distributions *P* and *Q*.**
- **Evaluate the relevance between two variables**

Information Gain

- **Based on Information Entropy**
  - **Information entropy (more specifically, Shannon entropy) is the expected value (average) of the information contained in each message.**

Information Entropy Calculation

If we have a set with k different values in it, we can calculate the entropy as follows:

$$entropy(Set) = I(Set) = -\sum_{i=1}^{k} P(value_i) \cdot \log_2(P(value_i))$$

Where $P(value_i)$ is the probability of getting the i^{th} value when randomly selecting one from the set.

So, for the set R = {a,a,a,b,b,b,b,b}

$$entropy(R) = I(R) = -\left[\underbrace{\left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right)}_{\text{a-values}} + \underbrace{\left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right)}_{\text{b-values}}\right]$$

Information Entropy Calculation

Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Small	Irregular	+
Yellow	Large	Irregular	+

Information Entropy Calculation

16 instances: 9 positive, 7 negative.

$$I(all\_data) = -\left[\left(\frac{9}{16}\right)\log_2\left(\frac{9}{16}\right) + \left(\frac{7}{16}\right)\log_2\left(\frac{7}{16}\right)\right]$$

This equals: 0.9836

This makes sense – it's almost a 50/50 split; so, the entropy should be close to 1.

Information Gain

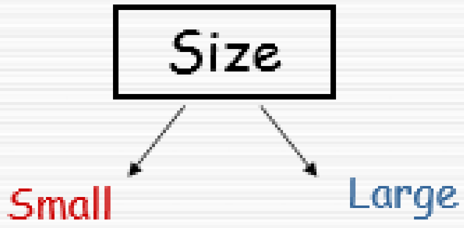
- **Definition**

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

***S*: the data set**

***A*: attribute**

Information Gain

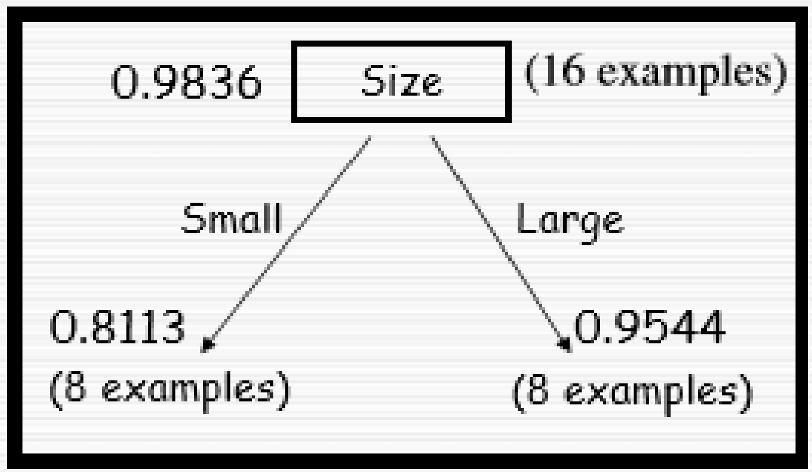


Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Small	Irregular	+

Color	Size	Shape	Edible?
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Irregular	+

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

Information Gain



Entropy of left child is 0.8113
I(size=small) = 0.8113

Entropy of right child is 0.9544
I(size=large) = 0.9544

$$I(S_{Size}) = (8/16)^{*}.8113 + (8/16)^{*}.9544 = .8828$$

Information Gain

$$G(attrib) = I(parent) - I(attrib)$$

We want to calculate the *information gain* (or entropy reduction). This is the reduction in 'uncertainty' when choosing our first branch as 'size'. We will represent information gain as "*G*."

$$G(size) = I(S) - I(S_{Size})$$

$$G(size) = 0.9836 - 0.8828$$

$$G(size) = 0.1008$$

Entropy of all data at parent node = I(parent) = 0.9836

Child's expected entropy for '**size**' split = I(size) = 0.8828

So, we have gained 0.1008 *bits* of information about the dataset by choosing 'size' as the first branch of our decision tree.

Information Gain

- **Statistical quantity measuring how well an attribute classifies the data.**
  - ✓ **Calculate the information gain for each attribute.**
  - ✓ **Choose attribute with greatest information gain.**

Information Gain

- **Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect.**
- **A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values.**

Outline

- Text Representation
- Feature Extraction
- Feature Selection
- **Summary**

Summary

- Representing raw data is the first step for building machine learning model.
- Feature extraction needs domain expert's help.
- Feature selection is necessary for improving performance and learning speed.
- Applications determine everything.

Thank you!

Q&A