Feature Representation and Selection

Marina Santini

The importance of good features

- Garbage in – garbage out

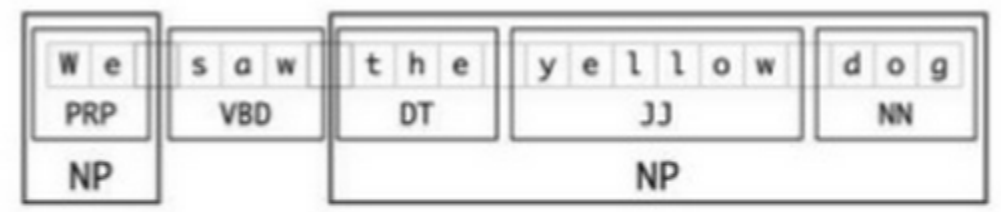
Bag of Words Representation

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

Parts Of Speech (PoS) representation



Segmentation and Labeling at both the Token and Chunk Levels

The Importance of Good Features

- Ex in Text Classification
  - BOW (Bag of words) (either counts or binary)
  - Phrases
  - n-Grams
  - Chunks
  - PoSs
  - PoS n-grams
  - etc.

ML success: Feature representation (aka feature engineering)

- The success of ML algorithms depens on how you present the data: you need great features that describe the structures inherent in your data:
  - Better features means flexibility
  - Better features means simpler models
  - Better features means better results

- However: The results you achieve are a factor of the model you choose, the data you have available and the features you prepared.

- That is, your results are dependent on many inter-dependent properties.

Feature Representation is a **knowledge representation** problem

- Transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data

- Question: what is the best representation of the sample data to learn a solution to your problem?

Practical Steps

[...] (tasks before here…)

- Select Data: Collect it together
- Preprocess Data: Format it, clean it, sample it so you can work with it.
- Transform Data: FEATURE REPRESENTATION happens here.
- Model Data: Create models, evaluate them and tune them.

[...] (tasks after here…)

Feature representation vs Feature Selection

- Feature representation is different from Attribute/Feature selection

Irrelevant and Reduntant Features

- Not all features have equal importance.
- Those attributes that are irrelevant to the problem need to be removed.
- Feature selection addresses those problems by automatically selecting a subset that are most useful to the problem.

The end