Discriminative Model Features

Making features from text for discriminative NLP models

Christopher Manning

Features

- In these slides and most maxent work: *features* f are elementary pieces of evidence that link aspects of what we observe *d* with a category *c* that we want to predict
- A feature is a function with a bounded real value: $f\colon C \times D \to \mathbb{R}$

Features

- In these slides and most maxent work: *features* f are elementary pieces of evidence that link aspects of what we observe d with a category c that we want to predict
- A feature is a function with a bounded real value

Example features

- $f_1(c, d) \equiv [c = \text{LOCATION} \wedge w_{-1} = \text{“in”} \wedge \text{isCapitalized}(w)]$
- $f_2(c, d) \equiv [c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$
- $f_3(c, d) \equiv [c = \text{DRUG} \wedge \text{ends}(w, \text{“c”})]$

LOCATION *in Arcadia* | LOCATION *in Québec* | DRUG *taking Zantac* | PERSON *saw Sue*

- Models will assign to each feature a *weight:*
  - A positive weight votes that this configuration is likely correct
  - A negative weight votes that this configuration is likely incorrect

Example features

- $f_1(c, d) \equiv [c = \text{LOCATION} \wedge w_{-1} = \text{“in”} \wedge \text{isCapitalized}(w)]$
- $f_2(c, d) \equiv [c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$
- $f_3(c, d) \equiv [c = \text{DRUG} \wedge \text{ends}(w, \text{“c”})]$

LOCATION	LOCATION	DRUG	PERSON
in Arcadia	*in Québec*	*taking Zantac*	*saw Sue*

- Models will assign to each feature a *weight:*
  - A positive weight votes that this configuration is likely correct
  - A negative weight votes that this configuration is likely incorrect

Feature Expectations

- We will crucially make use of two *expectations*
  - actual or predicted counts of a feature firing:

- Empirical count (expectation) of a feature:

$$\text{empirical } E(f_i) = \sum_{(c,d)\in \text{observed}(C,D)} f_i(c,d)$$

- Model expectation of a feature:

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

Features

- In NLP uses, usually a feature specifies (1) an indicator function – a yes/no boolean matching function – of properties of the input and (2) a particular class
  - $f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$ [Value is 0 or 1]
  - They pick out a data subset and suggest a label for it.
- We will say that $\Phi(d)$ is a feature of the data d, when, for each c_j, the conjunction $\Phi(d) \wedge c = c_j$ is a feature of the data-class pair (c, d)

Features

- In NLP uses, usually a feature specifies
  1. an indicator function – a yes/no boolean matching function – of properties of the input and
  2. a particular class

$$f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$$

[Value is 0 or 1]

- Each feature picks out a data subset and suggests a label for it

Christopher Manning

Feature-Based Models

- The decision about a data point is based only on the features active at that point.

Text Categorization

Data
BUSINESS: Stocks hit a yearly low …

Label: BUSINESS
Features
{…, stocks, hit, a, yearly, low, …}

Word-Sense Disambiguation

Data
… to restructure bank:MONEY debt.

Label: MONEY
Features
{…, w_{-1}=restructure, w_{+1}=debt, L=12, …}

POS Tagging

Data
DT JJ NN …
The previous fall …

Label: NN
Features
{w=fall, t_{-1}=JJ w_{-1}=previous}

Example: Text Categorization

(Zhang and Oles 2001)

- Features are presence of each word in a document and the document class (they do feature selection to use reliable indicator words)
- Tests on classic Reuters data set (and others)
  - Naïve Bayes: 77.0% F_1
  - Linear regression: 86.0%
  - Logistic regression: 86.4%
  - Support vector machine: 86.5%
- Paper emphasizes the importance of *regularization* (smoothing) for successful use of discriminative methods (not used in much early NLP/IR work)

Other Maxent Classifier Examples

- You can use a maxent classifier whenever you want to assign data points to one of a number of classes:
  - Sentence boundary detection (Mikheev 2000)
    - Is a period end of sentence or abbreviation?
  - Sentiment analysis (Pang and Lee 2002)
    - Word unigrams, bigrams, POS counts, …
  - PP attachment (Ratnaparkhi 1998)
    - Attach to verb or noun? Features of head noun, preposition, etc.
  - Parsing decisions in general (Ratnaparkhi 1997; Johnson et al. 1999, etc.)

Discriminative Model Features

Making features from text for discriminative NLP models

Christopher Manning