CSE528

Natural Language Processing

Venue:ADB-405 Topic: Text Classification

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

Is this spam?

From: "" <takworlld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the

methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===

Click here to order: http://www.wholesaledaily.com/sales/nmd.htm

===

Classification

Given:

- A description of an instance, $x \in X$, where X is the *instance language* or *instance space*.
  - Issue: how to represent text documents.
- A fixed set of categories:

  $C = \{c_1, c_2, \ldots, c_n\}$

Determine:

- The category of x: $c(x) \in C$, where $c(x)$ is a *categorization function* whose domain is X and whose range is C.
  - We want to know how to build categorization functions ("classifiers").

Examples

Labels are most often topics such as Yahoo-categories

e.g., "finance," "sports," "news>world>asia>business"

Labels may be genres

e.g., "editorials" "movie-reviews" "news“

Labels may be opinion

e.g., “like”, “hate”, “neutral”

Labels may be domain-specific binary

e.g., “spam” : “not-spam”, e.g., “contains adult language” :“doesn’t”

Classification Methods

Manual classification

- Used by Yahoo!, Looksmart, about.com, Medline
- Very accurate when job is done by experts
- Consistent when the problem size and team is small
- Difficult and expensive to scale

Automatic document classification

- Hand-coded rule-based systems
- E.g., assign category if document contains a given boolean combination of words
- Accuracy is often very high if a rule has been carefully refined over time by an expert
- Building and maintaining these rules is expensive

Classification Methods

Supervised learning of a document-label assignment function

- Many systems partly rely on machine learning
  - k-Nearest Neighbors (simple, powerful)
  - Naive Bayes (simple, common method)
  - Support-vector machines (new, more powerful)
  - Requires hand-classified training data
  - But data can be built up (and refined) by amateurs

Note that many commercial systems use a mixture of methods

Bayesian Methods

- Learning and classification methods based on probability theory.
- Bayes theorem plays a critical role in probabilistic learning and classification.
- Build a *generative model* that approximates how data is produced
- Uses *prior* probability of each category given no information about an item.
- Categorization produces a *posterior* probability distribution over the possible categories given a description of an item.

Bayes' Rule

$$P(C, X) = P(C \mid X)P(X) = P(X \mid C)P(C)$$

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Naive Bayes Classifiers

Task: Classify a new instance D based on a tuple of attribute values into one of the classes $c_j \in C$

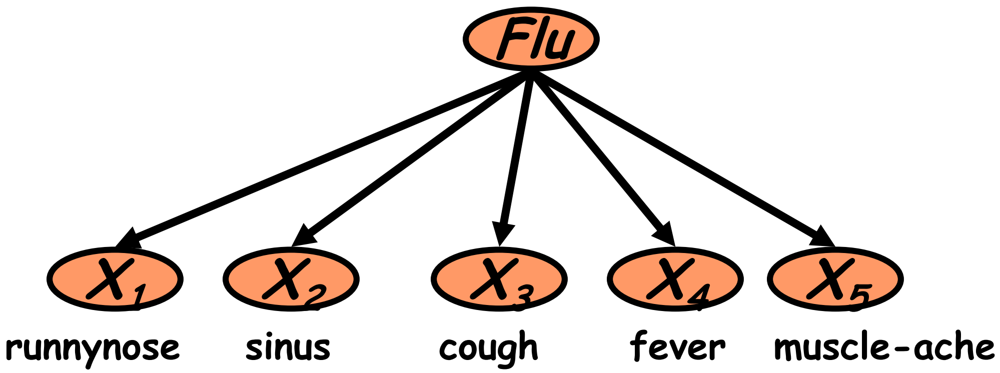
$$D = \langle x_1, x_2, \ldots, x_n \rangle$$

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c_j \in C}{\operatorname{argmax}}\ \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}}\ P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

The Naïve Bayes Classifier



Conditional Independence Assumption: features are independent of each other given the class

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

Learning the Model

First attempt: maximum likelihood estimates

- Simply use the frequencies in the data

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

Smoothing to Avoid Over fitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

# of values of X_i

Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

Calculate required $P(c_j)$ and $P(x_k \mid c_j)$ terms

- For each c_j in C do
  - $docs_j \leftarrow$ subset of documents for which the target class is c_j
  - $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$
  - $Text_j \leftarrow$ single document containing all $docs$
  - for each word x_k in *Vocabulary*
    - $n_k \leftarrow$ number of occurrences of x_k in $Text_j$
    - $$P(x_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Example

Training:

Document Name	Key Words						Class Name
	Kill	Bomb	Kidnap	Music	Movie	TV	
Doc1	2	1	3	0	0	1	Terrorism
Doc2	1	1	1	0	0	0	Terrorism
Doc3	1	1	2	0	1	0	Terrorism
Doc4	0	1	0	2	1	1	Entertainment
Doc5	0	0	1	1	1	0	Entertainment
Doc6	0	0	0	2	2	0	Entertainment

Testing:

Document Name	Key Words						Class Name
	Kill	Bomb	Kidnap	Music	Movie	TV	
Doc7	2	1	2	0	0	1	?

Example

$\lvert V \rvert$	C	$P(C_i)$	n_i	P(Kill / C_i)	P(Bomb / C_i)	P(Kidnap / C_i)	P(Music/ C_i)	P(Movie / C_i)	P(TV / C_i)
6	T	0.5	15	0.2380	0.1904	0.3333	0.0476	0.09523	0.09253
	E	0.5	12	0.0555	0.1111	0.1111	0.3333	0.2777	0.1111

|V| -> number of Vocabularies n_i -> total no 'of Documents

$P(C_i)$ -> no' of Documents in Class / no' of all Documents

$$P(\text{Kill} / T) = \frac{(2 + 1 + 1) + 1}{15 + |V|} = \frac{5}{21}$$

P(T / W) = P(T) * P(Kill / T) * P(Bomb / T) * P(Kidnap / T) * P(Music/ T) * P(Movie / T) * P(TV / T)

P(E/ W) = P(E) * P(Kill / E) * P(Bomb / E) * P(Kidnap / E) * P(Music/ E) * P(Movie / E) * P(TV / E)

Example

$P(T/W) = 0.5 * (0.2380)^2 * (0.1904)^1 * (0.3333)^2 * (0.0476)^0 * (0.09523)^0 * (0.09523)^1 = \mathbf{5.7047 \times 10^{-5}}$

$P(E/W) = 0.5 * (0.0555)^2 * (0.1111)^1 * (0.1111)^2 * (0.3333)^0 * (0.27777)^0 * (0.1111)^1 = \mathbf{2.3456 \times 10^{-5}}$

Since P(T/ W) has higher values therefore Document7 is classified into **Terrorism** Class

END