Topic: Text Classification

# Is this spam?

From: "" [<takworlld@hotmail.com](mailto: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**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,…, *c*n}

### Determine:

* The category of *x*: *c*(*x*)*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* |

*X* )*P*( *X* ) 

*P*( *X*

| *C*)*P*(*C*)

*P*(*C* |

*X* ) 

*P*( *X* | *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 *cj*  *C*

*D*  *x*1, *x*2 ,…, *xn*

*cMAP*

####  argmax

*c j* *C*

*P*(*cj*

| *x*1, *x*2 ,…, *xn* )

####  argmax

*P*(*x*1, *x*2 ,…, *xn* | *c j* )*P*(*cj* )

*c j* *C*

*P*(*x*1, *x*2 ,…, *xn* )

####  argmax

*c j* *C*

*P*(*x*1, *x*2 ,…, *xn*

| *cj* )*P*(*cj* )

# The Naïve Bayes Classifier

**runnynose**

***Flu***

***X1***

***X2***

***X3***

***X4***

***X5***

**sinus cough**

**fever**

**muscle-ache**

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

*P*(*X*1,…, *X*5

| *C*) 

*P*(*X*1

| *C*)  *P*(*X* 2

| *C*)  *P*(*X*5

| *C*)

# Learning the Model

## First attempt: maximum likelihood estimates

* Simply use the frequencies in the data

*P*ˆ(*x*

| *c* ) 

*N* ( *Xi*

 *xi* ,*C*

 *c j* )

*i j N* (*C*  *c* )

*j*

## Smoothing to Avoid Over fitting

*P*ˆ(*x*

| *c* ) 

*N* ( *Xi*

 *xi* , *C*

 *c j* ) 1

*i j N* (*C*  *c* )  *k*

*j*

# of values of *Xi*

Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

Calculate required *P*(*cj*) and *P*(*xk | cj*) terms

* For each *cj* in *C* do
  + *docsj*  subset of documents for which the target class is *cj*

*P*(*c* ) 



| *docs j* |

*j* | total# documents |

* *Textj*  single document containing all *docs*
* for each word *xk* in *Vocabulary*
  + *nk*  number of occurrences of *xk* in *Textj*
  + *P*(*x*

| *c* ) 

*nk*  

*k j n*   | *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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **|V|** | **C** | **P(Ci)** | **ni** | **P(Kill / Ci)** | **P(Bomb / Ci)** | **P(Kidnap / Ci )** | **P(Music/ Ci)** | **P(Movie / Ci)** | **P(TV / Ci)** |
| 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 ni -> total no 'of Documents P(Ci) -> no’ of Documents in Class / no’ of all Documents

P(Kill / T) = (2 + 1 + 1) +1 = 5

15 + |V| 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 = 5.7047 X 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 = 2.3456 X 10 -5**

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

