مبانی بازیابی اطلاعات و جستجوی وب

Text Classification & Naive Bayes -۱۳

some slides by: Hinrich Schütze and Christina Lioma

Outline

1. Text classification
2. Naive Bayes

A text classification task: Email spam filtering

```
From: ''' <takworlld@hotmail.com>
Subject: real estate is the only way...
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 Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
=================================================
```

How would you write a program that would automatically detect and delete this type of message?

Formal definition of TC: Training

Given:

- A document space X
  - Documents are represented in this space – typically some type of high-dimensional space.
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., relevant vs. nonrelevant).
- A training set D of labeled documents with each labeled document $\langle d, c \rangle \in X \times C$

Using a learning method or learning algorithm, we then wish to learn a classifier ϒ that maps documents to classes:

$$\Upsilon : X \rightarrow C$$

Formal definition of TC: Application/Testing

Given: a description $d \in X$ of a document Determine: $\Upsilon(d) \in C$, that is, the class that is most appropriate for d

6

Examples of how search engines use classification

- Language identification (classes: English vs. French etc.)
- The automatic detection of spam pages (spam vs. nonspam)
- Topic-specific or *vertical* search – restrict search to a "vertical" like "related to health" (relevant to vertical vs. not)
- Standing queries (e.g., Google Alerts)
- Sentiment detection: is a movie or product review positive or negative (positive vs. negative)

Classification methods: 1. Manual

- Manual classification was used by Yahoo in the beginning of the web.
- Very accurate if job is done by experts
- Consistent when the problem size and team is small
- Scaling manual classification is difficult and expensive.
- → We need automatic methods for classification.

8

Classification methods: 2. Rule-based

- Our Google Alerts example was rule-based classification.
- Often: Boolean combinations (as in Google Alerts)
- Accuracy is very high if a rule has been carefully refined over time by a subject expert.
- Building and maintaining rule-based classification systems is cumbersome and expensive.

9

Classification methods: 3. Statistical/Probabilistic

- This was our definition of the classification problem – text classification as a learning problem
- (i) Supervised learning of a the classification function ϒ and

  (ii) its application to classifying new documents
- But this manual classification can be done by non-experts.

تئوری بیز :تعریف مفاهیم اولیه

- فرض کنید که کلاسهای C و مجموعه مثالهای آموزش D موجود باشند. مقادیر احتمال زیر را تعریف میکنیم:

1. P(c)=احتمال اولیه ای که کلاس c قبل از مشاهده سند d داشته است (*prior probablity)* اگر چنین احتمالی موجود نباشد میتوان به تمامی فرضیه ها احتمال یکسانی نسبت داد.
2. P(d)=احتمال اولیه ای که سند d مشاهده خواهد شد.
3. P(d|c)=احتمال مشاهده سند d به فرض آنکه کلاس c صادق باشد.

- در رده بندی علاقه مند به دانستن P(c|d) یعنی احتمال اینکه با مشاهده سند dکلاس c صادق باشد، هستیم. این رابطه احتمال ثانویه (posterior probablity)نامیده میشود.
- توجه شود که احتمال اولیه مستقل از داده آموزشی است ولی احتمال ثانویه تاثیر داده آموزشی را منعکس میکند.

تئوری بیز

- سنگ بنای یادگیری بیزی را تئوری بیز تشکیل میدهد .این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه میدهد:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Prior probability

Posterior probability

Evidence

رده بند بیز

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

- برای محاسبه کلاس یک نمونه نیازی به محاسبه مخرج کسر نیست زیرا:

$$\underset{c}{argmax}\, P(c|d) = \underset{c}{argmax}\, \frac{P(d|c)P(c)}{P(d)}$$

$$= \underset{c}{argmax}\, P(d|c)P(c)$$

Outline

1. Text classification
2. Naive Bayes

فرض بیز ساده

- هدف ما مدلسازی p(d|c) می‌باشد. اما اگر به طور مثال ۵۰۰۰۰ کلمه داشته باشیم، تعداد پارامترها بسیار زیاد خواهد بود
- برای تخفیف این شرایط، فرض بیز ساده را انجام می‌دهیم: با داشتن c، ویژگی‌های ورودی t_i ها از یکدیگر مستقل هستند.
  - به طور مثال اگر فرض کنیم یک رایانامه، اسپم است (c=1)، دانش ما در مورد اینکه کلمه "buy" در پیام وجود دارد، تاثیری بر دانش ما در مورد اینکه کلمه "price" در پیام وجود دارد. ندارد.

Naive Bayes conditional independence assumption

To reduce the number of parameters to a manageable size, we make the Naive Bayes conditional independence assumption:

$$P(d|c) = P(\langle t_1, \ldots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

We assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(X_k = t_k \mid c)$.

The Naive Bayes classifier

- The Naive Bayes classifier is a probabilistic classifier.
- We compute the probability of a document *d* being in a class *c* as follows:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- n_d is the length of the document. (number of tokens)

18

Maximum a posteriori class

- Our goal in Naive Bayes classification is to find the "best" class.
- The best class is the most likely or maximum a posteriori (MAP) class c_{map}:

$$c_{map} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \ \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

Parameter estimation

- Estimate parameters $\hat{P}(c)$ and $\hat{P}(t_k|c)$ from train data: How?
- Prior:

$$\hat{P}(c) = \frac{N_c}{N}$$

- N_c : number of docs in class *c*; *N*: total number of docs
- Conditional probabilities:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- T_{ct} is the number of tokens of *t* in training documents from class *c* (includes multiple occurrences)

20

To avoid zeros: Add-one smoothing

- Before:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- Now: Add one to each count to avoid zeros:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

- B is the number of different words (in this case the size of the vocabulary: $|V| = M$)

21

Exercise

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

- Estimate parameters of Naive Bayes classifier
- Classify test document

22

Example: Parameter estimates

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$ Conditional probabilities:

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

The denominators are (8 + 6) and (3 + 6) because the lengths of $text_c$ and $text_{\overline{c}}$ are 8 and 3, respectively, and because the constant B is 6 as the vocabulary consists of six terms.

23

Example: Classification

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$
$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Thus, the classifier assigns the test document to c = *China*. The reason for this classification decision is that the three occurrences of the positive indicator `CHINESE` in d_5 outweigh the occurrences of the two negative indicators `JAPAN` and `TOKYO`.

24

منابع

- فصل سیزدهم کتاب An introduction to information retrieval