

# A Note on Naive Bayes Classifiers

Hiroya Takamura  
takamura@pi.titech.ac.jp

平成 20 年 1 月 25 日

## 概要

Naive Bayes Classifiers.

## 1 ナイブ・ベイズ分類器 ( Naive Bayes Classifiers )

今度は、確率が登場する。ラベル付事例から、ある事例が各クラスに属する確率を求め、その確率が最も大きいクラスのラベルを割り当てる。

ここでは、簡単のため、 $x_i$  は、0 もしくは 1 以上のどちらかの値を取るとする。  
まず、ベイズの定理をおさらいする。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

つまり、B が起こった時の A が起こる確率を知りたいのだが、直接は計算できないので、右辺の計算式を用いるというわけである。

これを文書分類で考えると、

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} \quad (2)$$

となる。つまり、左辺は「ある文書ベクトル  $\mathbf{x}$  が与えられた時の、クラス  $c$  の確率」である。これが直接わかれば、確率をもっとも大きいクラスを選べばよいので簡単である。ところが、当然これはわかっていない。で、右辺を使って計算する。

さて、実際 NLP への応用を考えると、この右辺、計算できるはずがない。問題は  $P(\mathbf{x}|c)$  である。これは、あるクラス  $c$  が起こった時の文書  $\mathbf{x}$  の生起確率である。限られたラベル付事例 ( クラス付文書 ) からあらゆる文書についてこのような確率が計算できるはずはないのである。  $\mathbf{x}$  の次元は普通単語数なので万単位であり、文書の種類は膨大な数になるからである。

そこで  $P(\mathbf{x}|c)$  を近似する。文書は、各単語が起こるか起こらないかでモデル化され、単語同士はお互いに独立である：

$$P(\mathbf{x}|c) \approx \prod_{w_i \in \mathbf{X}} P(w_i|c) \times \prod_{w_i \notin \mathbf{X}} (1 - P(w_i|c)) \quad (3)$$

としてみる。そうすると、この値の計算に必要なのは、 $P(x_i|c)$ 、つまり「クラスが与えられた時の各単語の生起確率」である。これならまあ計算できる。

$$P(x_i|c) = \frac{(\text{クラス } c \text{ であるような訓練文書のうち } x_i \text{ を含むものの数})}{(\text{クラス } c \text{ であるような訓練文書数})} \quad (4)$$

とでもしておけばよい (本当はよくない)<sup>1</sup>.

これを元の式に代入すると,

$$P(c|\mathbf{x}) = \frac{P(c) \prod_{w_i \in \mathbf{x}} P(w_i|c) \times \prod_{w_i \notin \mathbf{x}} (1 - P(w_i|c))}{P(\mathbf{x})} \quad (5)$$

となる. この値が最も大きくなるようなクラス  $c$  を,  $\mathbf{x}$  のクラスとすればよい.

が, 実際に上式の値の大小を比べるにあたって,  $P(\mathbf{x})$  は比較に影響を及ぼさない. よって,

$$P(c) \prod_{w_i \in \mathbf{x}} P(w_i|c) \times \prod_{w_i \notin \mathbf{x}} (1 - P(w_i|c)) \quad (6)$$

の大小でクラスを決定すればよいことになる. これが「ナイーブ・ベイズ分類器」である.

実は, ここで説明したモデルは, 「多変数ベルヌーイモデル (multivariate bernoulli model)」と呼ばれるナイーブベイズ分類器である. もう一つ有名なモデルに, 「多項モデル (multinomial model)」がある. 多変数ベルヌーイモデルでは, 各単語が起こったか起こらなかったかをモデル化しているのに対し, 多項モデルでは, 文書の各位置でどの単語が起こったかをモデル化している. 詳しくは, McCallum らの論文 [4] を参照してほしい. 多くの論文で, この二つのモデルの違いを理解してないと思われる記述が見られる.

多項モデルについても, 簡単に説明しておく. 多項モデルでは, 各単語の生起回数がモデルに組み込まれる:

$$P(\mathbf{x}|c) = P(|\mathbf{x}|) |\mathbf{x}|! \prod_i \frac{P(w_i|c)^{N(i, \mathbf{x})}}{N(i, \mathbf{x})!}, \quad (7)$$

ここで,  $P(|\mathbf{x}|)$  は長さ  $|\mathbf{x}|$  の文書が起こる確率を表し,  $N(i, \mathbf{x})$  は文書  $\mathbf{x}$  内での単語  $w_i$  の頻度を表す.

しかし,  $|\mathbf{x}|$  や  $N(i, \mathbf{x})$  は分類結果に影響しないので, モデル化の際は無視されることが多い. つまり,

$$P(\mathbf{x}|c) \propto \prod_i P(w_i|c)^{N(i, \mathbf{x})}, \quad (8)$$

で十分である.

パラメータ推定は,

$$P(w_i|c) = \frac{\sum_{\mathbf{x} \in c} N(i, \mathbf{x})}{\sum_i \sum_{\mathbf{x} \in c} N(i, \mathbf{x})} \quad (9)$$

とすればよい. これは, おおざっぱに言うと,

$$P(w|c) = \frac{\text{クラス } c \text{ に属する訓練文書全体での } w \text{ の出現回数}}{\text{クラス } c \text{ に属する訓練文書全体での全単語の出現回数}} \quad (10)$$

である. multivariate Bernoulli モデルのときの

$$P(w|c) = \frac{\text{クラス } c \text{ に属する訓練文書で } w \text{ を含む文書数}}{\text{クラス } c \text{ に属する訓練文書数}} \quad (11)$$

と異なることに注意してほしい.

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} \quad (12)$$

<sup>1</sup>訓練データにおいてクラス  $c$  内で  $x_i$  が起こらなかったとする. すると,  $x_i$  を含む全ての事例は  $P(\mathbf{x}|c) = 0$  となる.  $x_i$  以外の素性がクラス  $c$  であることをどんなに強く示唆していても.

$$= \frac{P(c)P(\mathbf{x}|c)}{\sum_c P(c)P(\mathbf{x}|c)} \quad (13)$$

$$= \frac{P(c)P(\mathbf{x}|c)}{\sum_c P(c)P(\mathbf{x}|c)} \quad (14)$$

$$= \frac{P(c)P(|\mathbf{x}|)|\mathbf{x}|! \prod_i \frac{P(w_i|c)^{N(i,\mathbf{X})}}{N(i,\mathbf{X})!}}{\sum_c P(c)P(|\mathbf{x}|)|\mathbf{x}|! \prod_i \frac{P(w_i|c)^{N(i,\mathbf{X})}}{N(i,\mathbf{X})!}} \quad (15)$$

$$= \frac{P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}}{\sum_c P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}} \quad (16)$$

## 1.1 Reversed Multinomial Naive Bayes

The Reversed Multinomial Naive Bayes Classifier is described by Juan and Ney [2].

$$P(c|\mathbf{x}) = \frac{P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}}{\sum_c P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}} \quad (17)$$

$$= \frac{P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}}{\sum_c P(c) \prod_i P(w_i|c)^{N(i,\mathbf{X})}} \quad (18)$$

$$= \frac{P(c) \prod_i \left( \frac{P(w_i)P(c|w_i)}{P(c)} \right)^{N(i,\mathbf{X})}}{\sum_c P(c) \prod_i \left( \frac{P(w_i)P(c|w_i)}{P(c)} \right)^{N(i,\mathbf{X})}} \quad (19)$$

$$= \frac{P(c) \prod_i \left( \frac{P(c|w_i)}{P(c)} \right)^{N(i,\mathbf{X})}}{\sum_c P(c) \prod_i \left( \frac{P(c|w_i)}{P(c)} \right)^{N(i,\mathbf{X})}} \quad (20)$$

They used uniform priors for category distributions. Then,

$$P(c|\mathbf{x}) = \frac{\prod_i P(c|w_i)^{N(i,\mathbf{X})}}{\sum_c \prod_i P(c|w_i)^{N(i,\mathbf{X})}} \quad (21)$$

## 1.2 Parameter Smoothing

As mentioned earlier, the Laplace estimation is often used for the parameter estimation in Naive Bayes classification [2] :

$$P(w_i|c) = \frac{N(i,c) + \delta}{\sum_i N(i,c) + \delta|V|}. \quad (22)$$

Many smoothing methods have been proposed in the literature of statistical language modelling. Those methods can be applied to our case.

- Absolute discounting with backing-off

$$P(w_i|c) = \begin{cases} \frac{N(i,c)-\delta}{\sum_i N(i,c)} & N(i,c) > 0 \\ \frac{1}{N_0} \frac{(|V|-N_0)\delta}{\sum_i N(i,c)} & N(i,c) = 0 \end{cases} \quad (23)$$

where  $N_0$  is the number of pairs  $\langle w_i, c \rangle$  that did not occur in the training data.

- Absolute discounting with interpolation

$$P(w_i|c) = \max\left\{0, \frac{N(i,c) - \delta}{\sum_i N(i,c)}\right\} + \frac{\sum_c N(i,c)}{\sum_i \sum_c N(i,c)} \frac{(|V| - N_0)\delta}{\sum_i N(i,c)} \quad (24)$$

$(|V| - N_0)\delta / \sum_i N(i,c)$  is the probability mass not assigned to the pairs  $\langle w_i, c \rangle$  whose  $N(i,c)$  is positive.

### 1.3 Poisson NB

### 1.4 Negative binomial NB

## 2 Obtaining more accurate probabilities with NB

### 2.1 Document length normalization

$$N'(i, \mathbf{x}) = L \frac{N(i, \mathbf{x})}{\sum_i N(i, \mathbf{x})} \quad (25)$$

where  $L$  is a constant.

The decision rule itself doesn't change with this normalization [2].

### 2.2 Histogram method (or Binning)

Zadrozny and Elkan [7] used *the histogram method* to obtain more accurate probabilities.

This method is applicable to only binary classification.

Suppose we have output values  $P(c|\mathbf{x})$  produced by a naive bayes classifier ( $c \in \{0, 1\}$ ). These values are usually not accurate [3].

We quote a sentence from the paper written by Zadrozny and Elkan:

We sort the training examples according to their scores and divide the sorted set into  $b$  subsets of equal size, called bins.

Then, the modified probability of a class given  $\mathbf{x}$  in the  $i$ -th bin  $B(i)$  is

$$\bar{P}(c|\mathbf{x}) = \frac{\{\mathbf{x}' \in B(i) \wedge \mathbf{x}' \in c \wedge \mathbf{x}' \in TR\}}{\{\mathbf{x}' \in B(i) \wedge \mathbf{x}' \in TR\}}. \quad (26)$$

The naive-bayes output values  $P(c|\mathbf{x})$  do not have to be probabilities. They can be scores as well.

Therefore, this method has also been applied to SVMs [1].

For multiclass case, see [8].

## 3 教師付き + 教師無し

さて、教師付き学習に用いられるラベル付事例集合は、人間が正しい答えを用意してあげなくてはならないので、大量に作るのは手間がかかる。しかし、教師無し学習で用いられるラベル無し事例集合は、比較的簡単に用意できる（例えばウェブからどさっと持ってくればよい）。

このような状況を踏まえて、ここでは教師無し学習を用いて教師付き学習を補強してあげる手法について考える。

### 3.1 Naive Bayes + EM

ナイーブ・ベイズとEMアルゴリズムを組み合わせる手法を紹介する。これは、[5]で提案された手法である。まず、ナイーブ・ベイズの式の、あとで不要になる部分を無視して、次のシンプルな式を得る：

$$P(\mathbf{x}|c, \theta) = \prod_i P(w_i|c)^{N(i, \mathbf{x})}. \quad (27)$$

すると、

$$P(\mathbf{x}|\theta) = \sum_c P(c) \prod_i P(w_i|c)^{N(i, \mathbf{x})}. \quad (28)$$

となる（EMの式の導出をわかりやすくするため、パラメータの $\theta$ を明示的に記述した）。

ここで $c$ を隠れ変数と見なす。E-ステップで計算される、隠れ変数の事後確率は、

$$P(c|\mathbf{x}, \bar{\theta}) = \frac{P(\mathbf{x}, c|\bar{\theta})}{P(\mathbf{x}|\bar{\theta})} \quad (29)$$

$$= \frac{P(\mathbf{x}, c|\bar{\theta})}{\sum_c P(\mathbf{x}, c|\bar{\theta})} \quad (30)$$

$$= \frac{P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})}{\sum_c P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})} \quad (31)$$

次に、Q関数を構成してみる：

$$Q(\theta|\bar{\theta}) = \sum_{\mathbf{x} \in D} \sum_c P(c|\mathbf{x}, \bar{\theta}) \log \left( P(c) \prod_i P(w_i|c)^{N(i, \mathbf{x})} \right). \quad (32)$$

さて、これを最大にするパラメータ $\theta$ を求めればよいのだが、単に偏微分しても今度はうまくいかない。というのは、 $\sum_c P(c|\bar{\theta}) = 1$ などの条件を明示的に組み込む必要があるからだ。さきほどのガウス分布の場合は、既に組み込まれていた（平均がいくつでも、確率分布としてvalidである）。

つまり、等式条件付き最大化問題を解くことになる。次のラグランジュ関数を用意する：

$$L(\theta|\bar{\theta}) = Q(\theta|\bar{\theta}) + \alpha \left( \sum_c P(c|\bar{\theta}) - 1 \right) + \sum_c \beta_c \left( \sum_w P(w|c, \bar{\theta}) - 1 \right) \quad (33)$$

$\alpha, \beta_c$ はラグランジュ乗数である。これを解くと、 $P(c|\bar{\theta}), P(w|c, \bar{\theta})$ の更新式が得られる。具体的な式はここではスキップする。

むしろ強調したいのは、どのようにしてラベル付き事例集合とラベル無し事例集合が使われているか、である。

それは、E-ステップでの計算にある。 $c$ が隠れ変数といっても、ラベル付き事例に関しては、隠れていない。よって、 $P(c|\mathbf{x}, \bar{\theta})$ の計算において、 $\mathbf{x}$ がラベル付きならその $c$ について値を1にし、それ以外のカテゴリについて0にする。ラベル無し事例については、上の計算式をそのまま使用する。このようにして、ラベル付き事例とラベル無し事例が効果的に利用されるのである。

[5]に載っている実験結果の一部を抜粋する。これは、20 Newsgroupと呼ばれるデータセットに対して、行われた実験であり、NB+EMに関しては、10000のラベル無し事例が使われている。

今回は、一つの例を紹介しただけだったが、ラベル無しデータの利用は、最近注目を集め、盛んに研究されている。

### 3.2 Calibration

The purpose of this method [6] is to obtain more accurate probabilities, when the estimated values are biased as a result of EM computation.

表 1: NB vs. NB+EM

tr	Naive Bayes	Naive Bayes + EM
20	20	35
300	52	66
5500	76	78

1. given unlabeled examples  $\mathbf{x}$ , transform the scores (probabilities)  $P(c|\mathbf{x})$  by sigmoid function  $f(t) = 1/(1 + \exp(-t))$ .
2. sort the transformed scores in descending order.
3. shift all the transformed scores so that the proportion of  $c$  be the same as that for training data.
4. transform the shifted scores back into probability values by  $F(T) = \log \frac{1-T}{T}$ .

Applicable only to binary classification.

## 参考文献

- [1] Joseph Drish. Obtaining calibrated probability estimates from support vector machines, 2001.
- [2] Alfons Juan and Hermann Ney. Reversing and smoothing the multinomial naive bayes text classifier. In *Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems*, pages 200–212, 2002.
- [3] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 4–15, 1998.
- [4] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.
- [5] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.
- [6] Yoshimasa Tsuruoka and Jun’ichi Tsujii. Training a naive bayes classifier via the em algorithm with a class distribution constraint. In Walter Daelemans and Miles Osborne, editors, *Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 127–134, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics. In association with HLT-NAACL 2003.
- [7] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, pages 609–616. Morgan Kaufmann, San Francisco, CA, 2001.
- [8] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates, 2002.