EM Algorithm

Nguyễn Phương Thái
Computer Science Department
Faculty of Information Technology, VNU UET

Outline

- Generative models
  - Refer to Prof. Ho Tu Bao's slides 25-26
- Naïve Bayes
- EM

Some key concepts in statistical machine learning

Generative model vs. discriminative model

Generative model

- Mô hình về quan hệ của **tất cả các biến**, mô tả việc các dữ liệu được ngẫu nhiên sinh ra trong mối liên quan với một số biến ẩn.
- Học một phân bố xác suất liên hợp (joint probability distribution) của các biến quan sát được và biến đích

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

- Tiêu biểu cho bài toán học với dữ liệu không nhãn (unlabeled data).

Discriminative model

- Mô hình về mối quan hệ phụ thuộc có điều kiện của **biến đích** với biến quan sát được (bỏ qua việc mô hình tường minh các biến quan sát được).
- Học một phân bố xác suất có điều kiện của biến đích khi có các biến quan sát

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(y_1, \ldots, y_n | x_1, \ldots, x_n)$$

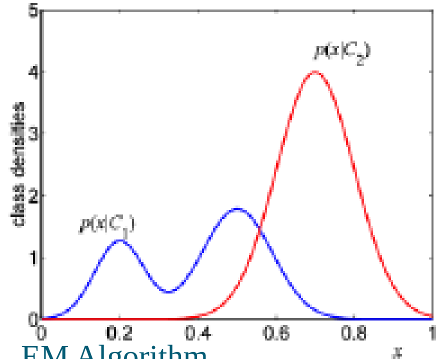
- Tiêu biểu cho bài toán học với dữ liệu có nhãn (labelled data).

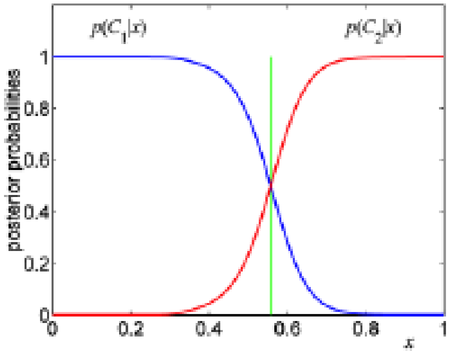
25

Some key concepts in statistical machine learning

Generative model vs. discriminative model

Generative model	Discriminative model
Học các hàm có dạng $p(\boldsymbol{x}\|\boldsymbol{y}), p(\boldsymbol{y})$. Ta ước lượng trực tiếp tham số $p(\boldsymbol{x}\|\boldsymbol{y}), p(\boldsymbol{y})$ từ dữ liệu huấn luyện, và từ đó dùng luật Bayes để tính $p(\boldsymbol{y}\|\boldsymbol{x})$. HMM, Markov random fields, Gaussian mixture models, Naïve Bayes, LDA, etc.	Học các hàm có dạng $p(\boldsymbol{y}\|\boldsymbol{x})$ Ước lượng tham số của $p(\boldsymbol{y}\|\boldsymbol{x})$ trực tiếp từ dữ liệu huấn luyện. SVM, logistic regression, neural networks, nearest neighbors, boosting, MEMM, conditional random fields, etc.





Naïve Bayes

- A simple but important probabilistic model for classification.
- First consider maximum-likelihood estimation in the case where the data is “fully observed”
- Then consider the expectation maximization (EM) algorithm for the case where the data is “partially observed”, in the sense that the labels for examples are missing.

Naïve Bayes

Assume we have some training set $(x^{(i)}, y^{(i)})$ for $i = 1 \ldots n$, where each $x^{(i)}$ is a vector, and each $y^{(i)}$ is in $\{1, 2, \ldots, k\}$.

Here k is an integer specifying the number of classes in the problem. This is a *multiclass* classification problem, where the task is to map each input vector x to a label y that can take any one of k possible values.

(For the special case of $k = 2$ we have a binary classification problem.)

We will assume throughout that each vector x is in the set $\{-1, +1\}^d$ for some integer d specifying the number of "features" in the model.

The Naive Bayes model is then derived as follows. We assume random variables Y and $X_1 \ldots X_d$ corresponding to the label y and the vector components $x_1, x_2, \ldots, x_d$. Our task will be to model the joint probability

$$P(Y = y, X_1 = x_1, X_2 = x_2, \ldots X_d = x_d)$$

for any label y paired with attribute values $x_1 \ldots x_d$. A key idea in the NB model is the following assumption:

$$\begin{aligned} & P(Y = y, X_1 = x_1, X_2 = x_2, \ldots X_d = x_d) \\ = \;& P(Y = y) \prod_{j=1}^{d} P(X_j = x_j | Y = y) \qquad (1) \end{aligned}$$

Following Eq. 1, the NB model has two types of parameters: $q(y)$ for $y \in \{1 \ldots k\}$, with

$$P(Y = y) = q(y)$$

and $q_j(x|y)$ for $j \in \{1 \ldots d\}$, $x \in \{-1, +1\}$, $y \in \{1 \ldots k\}$, with

$$P(X_j = x|Y = y) = q_j(x|y)$$

We then have

$$p(y, x_1 \ldots x_d) = q(y) \prod_{j=1}^{d} q_j(x_j|y)$$

The next section describes how the parameters can be estimated from training examples. Once the parameters have been estimated, given a new test example $\underline{x} = \langle x_1, x_2, \ldots, x_d \rangle$, the output of the NB classifier is

$$\arg \max_{y \in \{1 \ldots k\}} p(y, x_1 \ldots x_d) = \arg \max_{y \in \{1 \ldots k\}} \left(q(y) \prod_{j=1}^{d} q_j(x_j | y) \right)$$

Maximum Likelihood Estimation for NBMs

Given the training set $(x^{(i)}, y^{(i)})$ for $i = 1 \ldots n$, the log-likelihood function is

$$\begin{aligned}
L(\underline{\theta}) &= \sum_{i=1}^{n} \log p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{n} \log \left(q(y^{(i)}) \prod_{j=1}^{d} q_j(x_j^{(i)} | y^{(i)}) \right) \\
&= \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \log \left(\prod_{j=1}^{d} q_j(x_j^{(i)} | y^{(i)}) \right) \\
&= \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log q_j(x_j^{(i)} | y^{(i)}) \qquad (4)
\end{aligned}$$

Definition 2 (ML Estimates for Naive Bayes Models) *Assume a training set* $(x^{(i)}, y^{(i)})$ *for* $i \in \{1 \dots n\}$. *The maximum-likelihood estimates are then the parameter values* $q(y)$ *for* $y \in \{1 \dots k\}$, $q_j(x|y)$ *for* $j \in \{1 \dots d\}$, $y \in \{1 \dots k\}$, $x \in \{-1, +1\}$ *that maximize*

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log q_j(x_j^{(i)}|y^{(i)})$$

subject to the following constraints:

1. $q(y) \geq 0$ *for all* $y \in \{1 \dots k\}$. $\sum_{y=1}^{k} q(y) = 1$.
2. *For all* y, j, x, $q_j(x|y) \geq 0$. *For all* $y \in \{1 \dots k\}$, *for all* $j \in \{1 \dots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j(x|y) = 1$$

Theorem 1 *The ML estimates for Naive Bayes models (see definition 2) take the form*

$$q(y) = \frac{\sum_{i=1}^{n}[[y^{(i)} = y]]}{n} = \frac{count(y)}{n}$$

and

$$q_j(x|y) = \frac{\sum_{i=1}^{n}[[y^{(i)} = y \text{ and } x_j^{(i)} = x]]}{\sum_{i=1}^{n}[[y^{(i)} = y]]} = \frac{count_j(x|y)}{count(y)}$$

I.e., they take the form given in Eqs. 2 and 3.

ML Problem for NB with Missing Labels

We now describe the parameter estimation method for Naive Bayes when the labels $y^{(i)}$ for $i \in \{1 \dots n\}$ are missing. The first key insight is that for any example $\underline{x}$, the probability of that example under a NB model can be calculated by marginalizing out the labels:

$$p(\underline{x}) = \sum_{y=1}^{k} p(\underline{x}, y) = \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j|y) \right)$$

Given the training set $(x^{(i)})$ for $i = 1 \ldots n$, the log-likelihood function (we again use $\underline{\theta}$ to refer to the full set of parameters in the model) is

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log p(x^{(i)})$$

$$= \sum_{i=1}^{n} \log \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j^{(i)}|y) \right)$$

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log \left(q(y^{(i)}) \prod_{j=1}^{d} q_j(x_j^{(i)}|y^{(i)}) \right)$$

Definition 4 (ML Estimates for Naive Bayes Models with Missing Labels) *Assume a training set* $(x^{(i)})$ *for* $i \in \{1 \dots n\}$. *The maximum-likelihood estimates are then the parameter values* $q(y)$ *for* $y \in \{1 \dots k\}$, $q_j(x|y)$ *for* $j\{1 \dots d\}$, $y \in \{1 \dots k\}$, $x \in \{-1, +1\}$ *that maximize*

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j^{(i)}|y) \right) \qquad (10)$$

subject to the following constraints:

1. $q(y) \geq 0$ *for all* $y \in \{1 \dots k\}$. $\sum_{y=1}^{k} q(y) = 1$.

2. *For all* y, j, x, $q_j(x|y) \geq 0$. *For all* $y \in \{1 \dots k\}$, *for all* $j \in \{1 \dots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j(x|y) = 1$$

EM Algorithm for NBMs

Inputs: An integer k specifying the number of classes. Training examples $(x^{(i)})$ for $i = 1 \ldots n$ where each $x^{(i)} \in \{-1, +1\}^d$. A parameter T specifying the number of iterations of the algorithm.

Initialization: Set $q^0(y)$ and $q_j^0(x|y)$ to some initial values (e.g., random values) satisfying the constraints

- $q^0(y) \geq 0$ for all $y \in \{1 \ldots k\}$. $\sum_{y=1}^{k} q^0(y) = 1$.
- For all y, j, x, $q_j^0(x|y) \geq 0$. For all $y \in \{1 \ldots k\}$, for all $j \in \{1 \ldots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j^0(x|y) = 1$$

EM Algorithm for NBMs (cont)

Algorithm:
For $t = 1 \ldots T$

1. For $i = 1 \ldots n$, for $y = 1 \ldots k$, calculate

$$\delta(y|i) = p(y|\underline{x}^{(i)}; \underline{\theta}^{t-1}) = \frac{q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}{\sum_{y=1}^{k} q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}$$

2. Calculate the new parameter values:

$$q^t(y) = \frac{1}{n} \sum_{i=1}^{n} \delta(y|i) \qquad q_j^t(x|y) = \frac{\sum_{i:x_j^{(i)}=x} \delta(y|i)}{\sum_i \delta(y|i)}$$

Output: Parameter values $q^T(y)$ and $q^T(x|y)$.

EM Algorithm in General Form

Inputs: Sets $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{Y}$ is a finite set (e.g., $\mathcal{Y} = \{1, 2, \ldots k\}$ for some integer k). A model $p(x, y; \underline{\theta})$ that assigns a probability to each (x, y) such that $x \in \mathcal{X}$, $y \in \mathcal{Y}$, under parameters $\underline{\theta}$. A set of Ω of possible parameter values in the model. A training sample $x^{(i)}$ for $i \in \{1 \ldots n\}$, where each $x^{(i)} \in \mathcal{X}$. A parameter T specifying the number of iterations of the algorithm.

Initialization: Set $\underline{\theta}^0$ to some initial value in the set Ω (e.g., a random initial value under the constraint that $\underline{\theta} \in \Omega$).

Algorithm:

For $t = 1 \ldots T$

$$\underline{\theta}^t = \arg\max_{\underline{\theta} \in \Omega} Q(\underline{\theta}, \underline{\theta}^{t-1})$$

where

$$Q(\underline{\theta}, \underline{\theta}^{t-1}) = \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \delta(y|i) \log p(x^{(i)}, y; \underline{\theta})$$

and

$$\delta(y|i) = p(y|x^{(i)}; \underline{\theta}^{t-1}) = \frac{p(x^{(i)}, y; \underline{\theta}^{t-1})}{\sum_{y \in \mathcal{Y}} p(x^{(i)}, y; \underline{\theta}^{t-1})}$$

Output: Parameters $\underline{\theta}^T$.

Guarantees for the Algorithm

Theorem 4 *For any* $\underline{\theta}, \underline{\theta}^{t-1} \in \Omega$, $L(\underline{\theta}) - L(\underline{\theta}^{t-1}) \geq Q(\underline{\theta}, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1})$.

The quantity $L(\underline{\theta}) - L(\underline{\theta}^{t-1})$ is the amount of progress we make when moving from parameters $\underline{\theta}^{t-1}$ to $\underline{\theta}$. The theorem states that this quantity is lower-bounded by $Q(\underline{\theta}, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1})$.

Theorem 4 leads directly to the following theorem, which states that the likelihood is non-decreasing at each iteration:

Theorem 5 *For* $t = 1 \ldots T$, $L(\underline{\theta}^t) \geq L(\underline{\theta}^{t-1})$.

Proof: By the definitions in the algorithm, we have

$$\underline{\theta}^t = \arg\max_{\underline{\theta} \in \Omega} Q(\underline{\theta}, \underline{\theta}^{t-1})$$

It follows immediately that

$$Q(\underline{\theta}^t, \underline{\theta}^{t-1}) \geq Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1})$$

(because otherwise $\underline{\theta}^t$ would not be the arg max), and hence

$$Q(\underline{\theta}^t, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1}) \geq 0$$

But by theorem 4 we have

$$L(\underline{\theta}^t) - L(\underline{\theta}^{t-1}) \geq Q(\underline{\theta}^t, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1})$$

and hence $L(\underline{\theta}^t) - L(\underline{\theta}^{t-1}) \geq 0$. □

Proof of Theorem 4

$$
\begin{aligned}
L(\underline{\theta}) - L(\underline{\theta}^{t-1}) &= \sum_{i=1}^{n} \log \frac{\sum_y p(x^{(i)}, y; \underline{\theta})}{\sum_y p(x^{(i)}, y; \underline{\theta}^{t-1})} \\
&= \sum_{i=1}^{n} \log \sum_y \left(\frac{p(x^{(i)}, y; \underline{\theta})}{p(x^{(i)}; \underline{\theta}^{t-1})} \right) \\
&= \sum_{i=1}^{n} \log \sum_y \left(\frac{p(y|x^{(i)}; \underline{\theta}^{t-1}) \times p(x^{(i)}, y; \underline{\theta})}{p(y|x^{(i)}; \underline{\theta}^{t-1}) \times p(x^{(i)}; \underline{\theta}^{t-1})} \right) \qquad (13) \\
&= \sum_{i=1}^{n} \log \sum_y \left(\frac{p(y|x^{(i)}; \underline{\theta}^{t-1}) \times p(x^{(i)}, y; \underline{\theta})}{p(x^{(i)}, y; \underline{\theta}^{t-1})} \right) \\
&\geq \sum_{i=1}^{n} \sum_y p(y|x^{(i)}; \underline{\theta}^{t-1}) \log \left(\frac{p(x^{(i)}, y; \underline{\theta})}{p(x^{(i)}, y; \underline{\theta}^{t-1})} \right) \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{y} p(y|x^{(i)};\underline{\theta}^{t-1})\log p(x^{(i)},y;\underline{\theta}) - \sum_{i=1}^{n}\sum_{y} p(y|x^{(i)};\underline{\theta}^{t-1})\log p(x^{(i)},y;\underline{\theta}^{t-1}) \\
&= Q(\underline{\theta},\underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1},\underline{\theta}^{t-1}) \qquad (15)
\end{aligned}
$$

Applications

- Applications
  - Machine translation (word alignment)
  - HMMs
  - PCFGs
  - …
- Limitations
  - Local optimum

Key Points

- A parameter estimation method
  - maximum likelihood
  - applicable to generative models in case of incomplete training data
  - local optimum
  - efficient in practice

Thank you!