MACHINE LEARNING

Cao Văn Chung
cvanchung@hus.edu.vn
Informatics Dept., MIM, HUS, VNU Hanoi

Naive Bayes Classifier

Statistical Classifier

- Xét bài toán phân loại thống kê với C phân lớp (class) $1, 2, \ldots, C$:
- Cho dữ liệu $\mathbf{x} \in \mathbb{R}^d$, tính xác suất để $\mathbf{x}$ thuộc về phân lớp $k \in \{1, 2, \ldots, C\}$

$$p(y = k|\mathbf{x}) \qquad \text{hoặc viết gọn} \quad p(k|\mathbf{x}). \tag{1}$$

  Tức tính xác suất để đầu ra là class k với điều kiện đầu vào là $\mathbf{x}$.
- Xác suất $p(k|\mathbf{x})$, nếu tính được, cho phép xác định khả năng để $\mathbf{x}$ rơi vào mỗi lớp k.
- Từ đó, trong các bài toán phân loại, ta sẽ dự đoán đầu ra y là phân lớp có xác suất lớn nhất

$$y = c^* = \arg \max_{k \in \{1,\ldots,C\}} p(k|\mathbf{x}). \tag{2}$$

Bayesian-based approach

- Thực tế biểu thức $p(k|\mathbf{x})$ trong (2) thường không tính được trực tiếp.
- Áp dụng quy tắc Bayes:

$$p(k|\mathbf{x}) = \frac{p(\mathbf{x}|k)p(k)}{p(\mathbf{x})} \propto p(\mathbf{x}|k)p(k).$$

Ở đây đẳng thức thứ nhất là quy tắc Bayes; đẳng thức thứ hai là do dữ liệu quan sát $\mathbf{x}$ không phụ thuộc vào lớp k.

- Do đó (2) trở thành

$$\mathbf{x} \quad \text{thuộc phân lớp} \qquad y = c^* = \arg \max_{k \in \{1,\dots,C\}} p(\mathbf{x}|k)p(k). \qquad (3)$$

Bayesian-based approach

Trong (3):

- $p(k)$ - được hiểu là xác suất để một mẫu dữ liệu bất kỳ rơi vào phân lớp k.
- Từ đó, $p(k)$ có thể được tính theo ước lượng hợp lý cực đại (Maximum Likelihood Estimation - MLE), tức là tỷ lệ số phần tử thuộc lớp c trên tổng số phần tử tập dữ liệu

$$p(k) = \frac{\left|\{x \in \mathbf{X} : \text{label}(x) = k\}\right|}{|\mathbf{X}|}.$$

Bayesian-based approach

Xác định $p(\mathbf{x}|k)$:

- $p(\mathbf{x}|k)$ là phân phối xác suất của dữ liệu bên trong lớp k (xác suất để tồn tại dữ liệu $\mathbf{x}$ với điều kiện nó thuộc lớp k).

Trường hợp dữ liệu 1 chiều ($d = 1$, $\mathbf{x}$ - biến đơn):

- $\mathbf{x}$ - Dạng phân loại (Category): Sử dụng xác suất tiên nghiệm:

$$p(\mathbf{x}|k) = \frac{\left|\left\{x_m \in \text{class} \quad k^{th} : x_m = \mathbf{x}\right\}\right|}{\left|\left\{x_m \in \text{class} \quad k^{th}\right\}\right|}$$

- $\mathbf{x}$ - Dạng giá trị số (Numeric): Bổ sung giả thiết $\mathbf{x}$ là mẫu quan sát từ đại lượng ngẫu nhiên $\mathbf{X}$ tuân theo phân phối xác suất nào đó $\Rightarrow$ Có thể sử dụng công thức phân phối đó ước lượng tiên nghiệm cho $p(\mathbf{x}|k)$.

Với $\mathbf{x}$ là đại lượng ngẫu nhiên nhiều chiều, thì việc xác định phân bố $p(\mathbf{x}|k)$ là không khả thi.

Naive Bayes Classifier

Giả thiết Naive (Naive's assumption)

- Giả thiết: $\mathbf{x} = (x_1, \ldots, x_d)^T \mathbb{R}^d$, với các thành phần x_i là độc lập với nhau trong mọi phân lớp k. Lúc đó

$$p(\mathbf{x}|k) = p(x_1, x_2, \ldots, x_d|k) = \prod_{i=1}^{d} p(x_i|k). \tag{4}$$

- Giả thiết về sự độc lập của các chiều dữ liệu được gọi là *Naive*. Như tên của nó, giả thiết này quá chặt và ít khi có với dữ liệu thực. Mặc dù vậy cứ áp dụng nó lại cho những kết quả ngoài mong đợi.
- Cách phân loại dữ liệu này gọi là *Naive Bayes Classifier* (*NBC*).
- Do giả thiết này dẫn đến tính toán đơn giản hơn rất nhiều, nên phương pháp này rất phù hợp với dữ liệu lớn - large scale.

Naive Bayes Classifier

- Là phương pháp Supervised learning cho bài toán **Phân loại - Classification**: Với $\mathbf{x} \in \mathbb{R}^d$ - đầu vào, xác định phân lớp (C phân lớp) đầu ra từ

$$y = c^* = \arg \max_{k \in \{1,\dots,C\}} p(k) \prod_{i=1}^{d} p(x_i|k). \tag{5}$$

- Dữ liệu huấn luyện (Training set) $(\mathbf{X}, \mathbf{Y})$:

$$\mathbf{X} = \left\{\mathbf{x}^n\right\}_{n=1}^{N}, \mathbf{x}^n = (x_1^n, \dots, x_d^n)^T \in \mathbb{R}^d; \quad \mathbf{Y} = \left\{y_n\right\}_{n=1}^{N}, \mathbf{y}_n \in \{1, 2, \dots, C\}$$

Naive Bayes Classifier

- **Xác suất** $p(k)$ trong (5): (Nhắc lại)
  Tính bằng MLE (hoặc MAP) dựa trên Training Set

$$p(k) = \frac{\left|\{x^m \in \mathbf{X} : \text{label}(x^m) = k\}\right|}{|\mathbf{X}|} = \frac{\left|\{\text{class} \quad k\}\right|}{N}. \tag{6}$$

  trong (6), $\mathbf{X}$ là tập huấn luyện, do đó biết phân lớp đầu ra $y_m = \text{label}(x^m)$ ứng với đầu vào x^m.

- **Dự đoán**: với một điểm dữ liệu mới $\mathbf{x}$, phân lớp sẽ được xác định theo (5)

Naive Bayes Classifier

- Chú ý trong (5), các $p(x_i|k) < 1$ và có thể rất nhỏ, nên khi số chiều d rất lớn, tích $\prod_{i=1}^{d} p(x_i|k)$ sẽ rất bé và do đó nhạy cảm với sai số.
- Do đó thay cho (5), ta dùng log Nepe của nó, phân lớp đầu ra của $\mathbf{x}$ là

$$y = c^* = \arg \max_{k \in \{1,\dots,C\}} = \log(p(k)) + \sum_{i=1}^{d} \log(p(x_i|k)). \tag{7}$$

- Do hàm log đồng biến nên kết quả (dự đoán phân lớp đầu ra) tính theo (5) hay (7) không đổi.

Naive Bayes Classifier

Tính $p(\mathbf{x_i}|k)$ - $i = 1, \ldots, d$**;** $k = 1, 2 \ldots C$:

Từ gợi ý với dữ liệu 1 chiều (biến đơn $d = 1$), cách tính $p(\mathbf{x_i}|k)$ phụ thuộc vào kiểu dữ liệu x_i.

- x_i - Category: Dữ liệu có kiểu phân loại (lựa chọn)
  - x_i - Multinomial (Nhiều lựa chọn): Phương pháp *Multinomial Naive Bayes*
  - x_i - Binary (02 lựa chọn): Phương pháp *Bernoulli Naive*
- x_i - Numeric: Phương pháp *Gaussian Naive Bayes*.

Multinomial Naive Bayes

x_i - **Multinomial:**
(Trường dữ liệu có kiểu nhiều lựa chọn, VD: nhóm máu; ngành học; ngày trong tuần...)

- Trường x_i là dữ liệu kiểu nhiều lựa chọn (Category - Multinomial), $p(\mathbf{x_i}|k)$ được tính

$$p(\mathbf{x}_i|k) = \frac{\left|\left\{x^m \in \text{class} \quad k^{th} : x_i^m = \mathbf{x}_i\right\}\right|}{\left|\left\{x^m \in \text{class} \quad k^{th}\right\}\right|} \tag{8}$$

ở đây xét $x^m \in \mathbf{X}$ - Dữ liệu huấn luyện.

Multinomial Naive Bayes

Kỹ thuật Laplace smoothing cho trường hợp x_i - Multinomial:

- Trong (8), nếu không tồn tại $x^m \in \text{class} \quad k^{th} : x_i^m = \mathbf{x}_i$ thì $p(\mathbf{x}|k) = 0$
- Điều này dẫn đến kết quả theo (8) có thể không phù hợp thực tế:
  - Nếu một thuộc tính của $\mathbf{x}$ không xuất hiện trong tập Training của phân lớp $k \Rightarrow$ Xác suất để $\mathbf{x}$ thuộc k bằng 0 - Cho dù các thuộc tính khác của của $\mathbf{x}$ thuộc về k với xác suất rất cao!
- Để khắc phục, thay cho (8), ta sử dụng công thức

$$p(\mathbf{x}_i|k) \approx \hat{\lambda}_{ki} = \frac{\left|\left\{x^m \in \text{class} \quad k^{th} : x_i^m = \mathbf{x}_i\right\}\right| + \alpha}{\left|\left\{x^m \in \text{class} \quad k^{th}\right\}\right| + c.\alpha} \tag{9}$$

- trong (9): $\alpha > 0$ (tùy ý, thường lấy là 1); c là số lựa chọn của trường dữ liệu x_i, VD: $c = 4$ với x_i là nhóm máu (có 4 nhóm A, B, O, AB)...

Bernoulli Naive Bayes

x_i - **Binary (Category):**
Trường dữ liệu dạng 02 lựa chọn, VD: Giới tính; Có/không hút thuốc ...
Quy ước $\mathbf{x}_i \in \{0, 1\}$.

Áp dụng (8) cho trường hợp $x_i \in \{0, 1\}$, $p(\mathbf{x_i}|k)$ được tính

$$p(\mathbf{x}_i|k) = \begin{cases} \dfrac{\left|\left\{x^m \in \text{class} \quad k^{th} : x_i^m = 1\right\}\right|}{N_k} =: P_{i,k} & \text{nếu} \quad \mathbf{x}_i = 1 \\ \dfrac{\left|\left\{x^m \in \text{class} \quad k^{th} : x_i^m = 0\right\}\right|}{N_k} = 1 - P_{i,k} & \text{nếu} \quad \mathbf{x}_i = 0 \end{cases} \tag{10}$$

ở đây đặt $N_k = \left|\left\{x^m \in \text{class} \quad k^{th}\right\}\right|$ và $P_{i,k} := \dfrac{\left|\left\{x^m \in \text{class} \quad k^{th} : x_i^m = 1\right\}\right|}{N_k}$.

Bernoulli Naive Bayes

Áp dụng kỹ thuật Laplace smoothing để đảm bảo $0 < P_{i,k} < 1$, lúc đó (10) có thể viết lại

$$p(\mathbf{x}_i|k) = \begin{cases} \left(P_{i,k}\right)^{\mathbf{x}_i} & \text{nếu} \quad \mathbf{x}_i = 1 \\ \left(1 - P_{i,k}\right)^{(1-\mathbf{x}_i)} & \text{nếu} \quad \mathbf{x}_i = 0 \end{cases}$$

Kết hợp hai trường hợp trên và áp dụng $a^0 = 1$ với $\forall a > 0$, ta có công thức tính $p(\mathbf{x}_i|k)$ cho trường hợp $\mathbf{x}_i \in \{0, 1\}$:

$$p(\mathbf{x}_i|k) = \left(P_{i,k}\right)^{\mathbf{x}_i} \left(1 - P_{i,k}\right)^{(1-\mathbf{x}_i)} \tag{11}$$

Công thức tính $p(\mathbf{x}_i|k)$ theo (11) gọi là **phương pháp Bernoulli**.

Gaussian Naive Bayes

$\mathbf{x}_i$ - **Numeric**: (Trường dữ liệu $\mathbf{x}_i$ là số thực, VD: tuổi, độ pH máu, điểm thi...)

- **Giả thiết**: Trong mỗi phân lớp k, dữ liệu x_i tuân theo phân bố chuẩn có kỳ vọng μ_{ki} và phương sai σ_{ki}^2: $\mathbf{x}_i \sim \mathcal{N}(\mu_{ki}, \sigma_{ki}^2)$.
- Lúc đó

$$p(x_i|k) = p(x_i|\mu_{ki}, \sigma_{ki}^2) \propto \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right). \tag{12}$$

- Trong đó bộ tham số $\theta = \{\mu_{ki}, \sigma_{ki}^2\}$ được xác định bằng Maximum Likelihood

$$(\mu_{ki}, \sigma_{ki}^2) = \arg\max_{\mu_{ki}, \sigma_{ki}^2} \prod_{n=1}^{N} p(x_i^{(n)}|\mu_{ki}, \sigma_{ki}^2). \tag{13}$$

Gaussian Naive Bayes

- Trong (12) ta sử dụng hàm mật độ (thay cho phân bố) vì bản chất ta cần xác định k cho $p(x_i|k)$ lớn nhất; đồng thời ta cần tìm tần suất xẩy ra x_i trong lớp thứ k.
- Giá trị tham số $\{\mu_{ki}, \sigma_{ki}^2\}$ được tính trên mẫu là tập Training theo công thức

$$\mu_{ki} = \frac{1}{N_k} \sum_{x^m \in k^{th}-class} x_i^m; \quad \sigma_{ki}^2 = \frac{1}{N_k} \sum_{x^m \in k^{th}-class} (x_i^m - \mu_{ki})^2 \qquad (14)$$

- Trong (14), $x^m \in \mathbf{X}$ - Tập huấn luyện (Training set); N_k - số mẫu thuộc class k.

Multinomial Naive Bayes' application

- Áp dụng mô hình Multinomial NB trong phân loại văn bản mà feature vectors được tính bằng *Bags of Words* (tự tìm hiểu các kỹ thuật trích xuất dữ liệu - *Feature Engineering*).
- Trong kỹ thuật này, mỗi văn bản được biểu diễn bởi một vector có độ dài d - là số từ trong từ điển. Dễ thấy số chiều sẽ rất lớn.
- Giá trị của thành phần (tọa độ) thứ i của mỗi vector chính là số lần từ thứ i (trong từ điển) xuất hiện trong văn bản.

Multinomial Naive Bayes' application

- Theo cách đặt trên $p(x_i|c)$ sẽ là tần suất xuất hiện từ thứ i trong toàn bộ các văn bản của lớp c. Giá trị này thường được tính theo công thức

$$\lambda_{ci} = p(x_i|c) = \frac{N_{ci}}{N_c}. \tag{15}$$

Ở đây

- N_{ci} là tổng số lần từ thứ i xuất hiện trong các văn bản thuộc lớp c. Nó được tính là tổng của tất cả các thành phần (tọa độ) thứ i của các điểm dữ liệu (feature vectors) trong phân lớp c.
- N_c là tổng số từ (kể cả lặp) xuất hiện trong phân lớp c. Tức là N_c bằng tổng độ dài tính theo từ của toàn bộ các văn bản thuộc vào lớp c.
- Có thể suy ra $N_c = \sum_{i=1}^{d} N_{ci}$, và do đó $\sum_{i=1}^{d} \lambda_{ci} = 1$.

Multinomial Naive Bayes

- Nhược điểm: Nếu một từ không xuất hiện trong phân lớp c thì $\lambda_{ci} = p(x_i|c) = 0$ dẫn tới trong (15) $p(\mathbf{x}|c) = 0$, dù các từ khác có tần suất lớn.
- Đặc điểm này sẽ dẫn đến kết quả không chính xác.
- Để tránh nhược điểm này, áp dụng kỹ thuật *Laplace smoothing*:

$$\hat{\lambda}_{ci} = \frac{N_{ci} + \alpha}{N_c + d\alpha}. \tag{16}$$

  - $\alpha > 0$ thường bằng 1, để tránh trường hợp tử số bằng 0.
  - Mẫu số được cộng với $d\alpha$ nên có thể đảm bảo tổng xác suất $\sum_{i=1}^{d} \hat{\lambda}_{ci} = 1$.
- Bây giờ mỗi lớp c sẽ được mô tả bằng một bộ d số dương có tổng bằng 1: $\hat{\lambda}_c = \{\hat{\lambda}_{c1}, \ldots, \hat{\lambda}_{cd}\}$.

Bernoulli Naive Bayes' application

- Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bẳng 0 hoặc 1.
- **Ví dụ**: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không.
- Trong trường hợp này, công thức tính các $p(x_i|c)$ như sau

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c)^{1-x_i}. \tag{17}$$

- $p(i|c)$ có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của lớp c.

Eg.1. Spam email - Classification

- Xét bài toán phân loại mail Spam (S) và Not Spam (N).
- Ta có bộ training data gồm E_1, E_2, E_3. Cần phân loại E_4.
- Bảng từ vựng: $[w_1, w_2, w_3, w_4, w_5, w_6, w_7]$.
- Ví dụ này có thể xử lý bằng Multinomial Naive Bayes hoặc Bernoulli Naive Bayes. Tuy nhiên ta sẽ sử dụng Multinomial Naive Bayes.
- Bảng thống kê số lần xuất hiện của từng từ trong từng email tương ứng có trong trang sau

Eg.1. Spam email - Classification

	Email	w_1	w_2	w_3	w_4	w_5	w_6	w_7	Label
Training data	E1	1	2	1	0	1	0	0	N
	E2	0	2	0	0	1	1	1	N
	E3	1	0	1	1	0	2	0	S
Test data	E4	1	0	0	0	0	0	1	?

- Tính các $p(k)$: Ta có $P(S) = \frac{1}{3}$, $P(N) = \frac{2}{3}$.
- Sử dụng Laplace Smoothing với $\alpha = 1$ ta tính được xác suất xuất hiện của từng từ trong văn bản như sau

Eg.1. Spam email - Classification

	Email	w_1	w_2	w_3	w_4	w_5	w_6	w_7
	E3	1	0	1	1	0	2	0
$P(w_i\|S)$	(trước Smoothing)	1/5	0/5	1/5	1/5	0/5	2/5	0/5
$P(w_i\|S)$	(sau Smoothing)	2/12	1/12	2/12	2/12	1/12	3/12	1/12

class = Not Spam (N)

	Email	w_1	w_2	w_3	w_4	w_5	w_6	w_7
	E1	1	2	1	0	1	0	0
	E2	0	2	0	0	1	1	1
Tổng		1	4	1	0	2	1	1
$P(w_i\|N)$	(trước Smoothing)	1/10	4/10	1/10	0/10	2/10	1/10	1/10
$P(w_i\|N)$	(sau Smoothing)	2/17	5/17	2/17	1/17	3/17	2/17	2/17

Eg.1. Spam email - Classification

Từ đó tính được

$$
\begin{aligned}
P(S|E_4) &\propto P(S)\prod_{i=1}^{7} P(w_i|S) \\
&\propto \frac{1}{3}\times(\frac{2}{12}\times\frac{1}{12}) \\
&\propto 0.0046
\end{aligned}
$$

$$
\begin{aligned}
P(N|E_4) &\propto P(N)\prod_{i=1}^{7} P(w_i|N) \\
&\propto \frac{2}{3}\times(\frac{2}{17}\times\frac{2}{17}) \\
&\propto 0.0092
\end{aligned}
$$

Eg.1. Spam email - Classification

Xác suất phân lớp tương ứng

$$P(S|E_4) = \frac{0.0046}{0.0046 + 0.0092} \approx 0.334$$
$$P(N|E_4) = \frac{0.0092}{0.0046 + 0.0092} \approx 0.666$$

Do đó ta phân loại E_4 là Not Spam (N).