MỘT SỐ THUẬT TOÁN PHÂN LOẠI VĂN BẢN

Lê Hồng Phương

<*phuonglh@gmail.com*>
Đại học Quốc gia Hà Nội
Trường Đại học Khoa học Tự nhiên
Viện Nghiên cứu Công nghệ FPT

6/2013

Nội dung

Nội dung

Nội dung

Bài toán phân loại văn bản

Bài toán

Cho $\mathbf{x}$ *là một văn bản. Biết* $\mathbf{x}$ *thuộc một trong các loại* $y \in \{1, 2, \ldots, K\}$. *Hãy tìm loại văn bản đúng nhất của* $\mathbf{x}$.

Ví dụ:

- Giả sử $\mathbf{x}$ là một bài báo do phóng viên viết, gửi đăng trên trang tin điện tử vnExpress. Biên tập viên cần quyết định xem $\mathbf{x}$ thuộc thể loại nào là thích hợp nhất: "*chính trị – xã hội*", "*quốc tế*", "*thể thao*"...
- Giả sử $\mathbf{x}$ là một văn bản ngắn có mục tiêu điều khiển tivi. Mỗi thể loại tương ứng với một hành động điều khiển: "*tắt*", "*bật*", "*chuyển kênh*",...:
  - $\mathbf{x}$ = "*hãy bật tivi*" $\Rightarrow y =$ "*bật*"
  - $\mathbf{x}$ = "*chuyển sang kênh HBO*" $\Rightarrow y =$ "*chuyển kênh*"

Bài toán phân loại văn bản

- Gọi $y = h_\theta(\mathbf{x})$ là hàm phân loại của $\mathbf{x}$ trong đó θ là tham số của hàm. Ta cần tìm $h_\theta(\cdot)$ có khả năng phân loại tốt.
- Để tìm h_θ, ta sử dụng *phương pháp học có hướng dẫn* từ dữ liệu mẫu:
  - Dữ liệu học gồm N mẫu: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$.
  - Hàm h_θ được xây dựng sao cho nó *khớp nhất* với dữ liệu huấn luyện này.

Bài toán phân loại văn bản

Mỗi văn bản $\mathbf{x}$ là một đối tượng cần phân loại, thông thường $\mathbf{x}$ được chuyển thành một biểu diễn véc-tơ thực D chiều:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D), \quad x_j \in \mathbb{R}$$

Các thành phần $x_j, j = 1, 2, \ldots, D$ được gọi là các *đặc trưng* hay *thuộc tính* của $\mathbf{x}$.

Nội dung

Mô hình xác suất

- Có nhiều phương pháp phân loại văn bản nhưng phương pháp phân loại cho kết quả tốt nhất hiện nay đều sử dụng các mô hình xác suất.
- Gọi $h_\theta(\mathbf{x}) = P(y|\,\mathbf{x};\theta)$ là mô hình xác suất có điều kiện dự báo các khả năng hay xác suất thuộc loại y của đối tượng $\mathbf{x}$.
- Đối tượng $\mathbf{x}$ sẽ được xếp vào loại có xác suất lớn nhất theo mô hình:

$$\widehat{y} = \underset{k=1,2,\ldots,K}{\arg\max}\ P(y = k|\,\mathbf{x};\theta)$$

- Chú ý rằng trong mô hình xác suất thì

$$\sum_{k=1,2,\ldots,K} P(y = k|\,\mathbf{x};\theta) = 1.$$

Nội dung

Nội dung

Mô hình Bayes đơn giản dạng nhị thức

- Giả sử $\mathbf{x}$ là một văn bản chứa các từ thuộc từ điển gồm D từ, đánh số từ 1 tới D.
- Khi đó ta có thể biểu diễn $\mathbf{x}$ bởi véc-tơ nhị phân

$$\mathbf{x} = (x_1, x_2, \ldots, x_D), \quad x_j \in \{0, 1\},$$

trong đó

$$x_j = \begin{cases} 1, & \text{nếu từ thứ } j \text{ xuất hiện trong } \mathbf{x} \\ 0, & \text{nếu từ thứ } j \text{ không xuất hiện trong } \mathbf{x}. \end{cases}$$

Mô hình Bayes đơn giản dạng nhị thức

Trong mô hình Bayes đơn giản (naive Bayes–NB), ta giả định các đặc trưng $x_j \in \{0, 1\}$ và độc lập nhau đối với từng loại y. Từ đó

$$\begin{aligned} P(\mathbf{x}, y; \theta) &= P(\mathbf{x}\,|y; \theta) P(y; \theta) \\ &= \prod_{j=1}^{D} P(x_j|y; \theta) P(y; \theta). \end{aligned}$$

Các tham số của mô hình:

$$\begin{aligned} \theta_k &= P(y = k), \forall k = 1, 2, \ldots, K \\ \theta_{j|k} &= P(x_j = 1|y = k), \forall j = 1, 2, \ldots, D; \forall k = 1, 2, \ldots, K \end{aligned}$$

Chú ý rằng $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$, nên mô hình có $(K - 1) + DK$ tham số.

Mô hình Bayes đơn giản dạng nhị thức

- Hàm hợp lí trên tập dữ liệu huấn luyện $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ là

$$L(\theta) = \prod_{i=1}^{N} P(\mathbf{x}_i, y_i) = \prod_{i=1}^{N} \left(\prod_{j=1}^{D} P(x_j|y;\theta) P(y;\theta) \right)$$
$$= \prod_{i=1}^{N} \left(\prod_{j=1}^{D} \theta_{j|k} \theta_k \right).$$

- Từ đó có ước lượng hợp lí cực đại:

$$\widehat{\theta}_k = \frac{\sum_{i=1}^{N} \delta(y_i = k)}{N},$$
$$\widehat{\theta}_{j|k} = \frac{\sum_{i=1}^{N} \delta(x_{ij} = 1, y_i = k)}{\sum_{i=1}^{N} \delta(y_i = k)},$$

trong đó $\delta(\cdot)$ là hàm chỉ số.

Mô hình Bayes đơn giản dạng đa thức

- Trong mô hình Bayes dạng đa thức, ta xét tần số xuất hiện của từng từ trong $\mathbf{x}$ thay vì chỉ xét từ đó có xuất hiện hay không như trong mô hình Bayes nhị phân.
- Tham số của mô hình:
  - θ_k là xác suất tiên nghiệm văn bản thuộc lớp k;
  - $\theta_{j|k}$ xác suất từ j xuất hiện trong lớp k.
- Gọi $f(k,j)$ là số lần từ j xuất hiện trong loại văn bản k. Khi đó ước lượng hợp lí cực đại của các tham số là

$$\widehat{\theta}_k = \frac{\sum_{i=1}^{N} \delta(y_i = k)}{N},$$

$$\widehat{\theta}_{j|k} = \frac{f(k,j)}{\sum_{j=1}^{D} f(k,j)}$$

Quy tắc phân loại

- Với mỗi đối tượng $\mathbf{x}$, ta phân nó vào loại y với

$$y := \widehat{k} = \arg\max_{k=1,2,\ldots,K} P(y = k|\,\mathbf{x})$$

$$= \arg\max_{k=1,2,\ldots,K} \prod_{j=1}^{D} \theta_{j|k}\theta_k.$$

- Nếu sử dụng hàm loga, ta có quy tắc phân loại tuyến tính:

$$y := \widehat{k} = \arg\max_{k=1,\ldots,K} \left(\sum_{j=1}^{D} \log \theta_{j|k} + \log \theta_k \right).$$

Quy tắc phân loại

Với mỗi văn bản $\mathbf{x}$, gọi $\mathcal{V}$ là tập từ thuộc $\mathbf{x}$. Thuật toán phân loại $\mathbf{x}$ trong mô hình Bayes đơn giản như sau:

Algorithm 1: Thuật toán phân loại Bayes đơn giản

Data: $\mathbf{x}, \theta_k, \theta_{j|k}, k = 1, 2, \ldots, K, j = 1, 2, \ldots, D$

for $k = 1, 2, \ldots, K$ **do**

    $s[k] \leftarrow \log \theta_k$;

    for $j \in \mathcal{V}$ **do**

        $s[k] \leftarrow s[k] + \log \theta_{j|k}$;

return $\arg\max_k s[k]$;

Làm trơn mô hình

- Ta cần làm trơn mô hình để xử lí trường hợp $\theta_{j|k} = 0$.
- Nếu $\theta_{j|k} = 0, \forall k = 1, 2, \ldots, K$ thì

$$P(\mathbf{x}) = \sum_{k=1}^{K} \left(\theta_k \prod_{j=1}^{D} \theta_{j|k} \right) = 0.$$

- Từ đó

$$P(y = k|\,\mathbf{x}) = \frac{0}{0}, \quad \forall k = 1, 2, \ldots, K$$

nên ta không thể phân loại $\mathbf{x}$.

Làm trơn mô hình

Ta sử dụng phương pháp làm trơn Laplace:

$$\widehat{\theta}_{j|k} = \frac{\widehat{\theta}_{j|k} + \alpha}{\widehat{\theta}_k + D \times \alpha}$$

trong đó α là hệ số làm trơn.

Nội dung

Mô hình Bernoulli

- Trong mô hình Bayes đơn giản ở trên, $\theta_{j|k}$ là *tần suất từ* hay *tần suất vị trí* trong các văn bản thuộc lớp k có chứa từ j.
- Mô hình Bernoulli sử dụng tham số này theo cách khác, là *tần suất văn bản* thuộc lớp k có chứa từ j.
- Như vậy, mô hình Bernoulli chỉ sử dụng thông tin từ j có xuất hiện trong văn bản lớp k hay không, không quan tâm từ đó xuất hiện bao nhiêu lần.

Mô hình Bernoulli

Gọi $f(k, j)$ là số lần văn bản thuộc loại k chứa từ j. Khi đó

$$\widehat{\theta}_{j|k} = \frac{f(k, j)}{\sum_{j=1}^{D} f(k, j)}.$$

Làm trơn mô hình:

$$\widehat{\theta}_{j|k} = \frac{\widehat{\theta}_{j|k} + \alpha}{\widehat{\theta}_k + D \times \alpha}$$

trong đó α là hệ số làm trơn.

Quy tắc phân loại

Với mỗi văn bản $\mathbf{x}$, gọi $\mathcal{V}$ là tập từ thuộc $\mathbf{x}$. Thuật toán phân loại $\mathbf{x}$ trong mô hình Bernoulli như sau:

Algorithm 2: Thuật toán phân loại Bernoulli

Data: $\mathbf{x}, \theta_k, \theta_{j|k}, k = 1, 2, \ldots, K, j = 1, 2, \ldots, D$

for $k = 1, 2, \ldots, K$ **do**
    $s[k] \leftarrow \log \theta_k$;
    for $j = 1, 2, \ldots, D$ **do**
        if $j \in \mathcal{V}$ **then**
            $s[k] \leftarrow s[k] + \log \theta_{j|k}$;
        else
            $s[k] \leftarrow s[k] + \log(1 - \theta_{j|k})$;

return $\arg\max_k s[k]$;

Nội dung

Mô hình TF-IDF

- Gọi $\text{tf}(j, \mathbf{x})$ là số lần từ j xuất hiện trong văn bản $\mathbf{x}$ và $\text{df}(j)$ là số văn bản có chứa từ j trong tập huấn luyện.
- Ta tính nghịch đảo của tần số văn bản chứa từ j như sau:

$$\text{idf}(j) = \log\left(\frac{N}{\text{df}(j)}\right).$$

- Về mặt trực quan, $\text{idf}(j)$ là nhỏ nếu từ j xuất hiện trong nhiều văn bản và là lớn nhất nếu nó chỉ xuất hiện trong một văn bản.
- Mỗi văn bản được biểu diễn dạng véc-tơ $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, trong đó

$$x_j = \text{tf}(j, \mathbf{x}) \times \text{idf}(j), \quad \forall j = 1, 2, \ldots, D.$$

Mô hình TF-IDF

- Tiếp theo, ta tính các tham số của mô hình:

$$\mathbf{c}_k = \sum_{i:y_i=k} \mathbf{x}_i, \quad \forall k = 1, 2, \ldots, K.$$

- Quy tắc phân loại cho văn bản $\mathbf{x}$ là:

$$\begin{aligned} y &= \arg\max_k \cos(\mathbf{x}, \mathbf{c}_k) \\ &= \arg\max_k \frac{\sum_{j=1}^{D} x_j \times c_{kj}}{\sqrt{\sum_{j=1}^{D} x_j^2} \times \sqrt{\sum_{j=1}^{D} c_{kj}^2}} \end{aligned}$$

Quy tắc phân loại

Với mỗi văn bản $\mathbf{x}$, gọi $\mathcal{V}$ là tập từ thuộc $\mathbf{x}$. Thuật toán phân loại $\mathbf{x}$ trong mô hình TF-IDF như sau:

Algorithm 3: Thuật toán phân loại TF-IDF

Data: $\mathbf{x}, \mathbf{c}_k, k = 1, 2, \ldots, K$
for $k = 1, 2, \ldots, K$ **do**
  $s[k] \leftarrow \cos(\mathbf{x}, \mathbf{c}_k)$;
return $\arg\max_k s[k]$;

Nội dung

Thiết kế

- Xem tài liệu thiết kế chi tiết các lớp trong gói phần mềm `com.fpt.nao.text`.
- Các lớp chính cài đặt các thuật toán phân loại:
  - `NBTextClassifier`
  - `BernoulliClassifier`
  - `TFIDFClassifier`
- Lớp `TextClassifierTester` minh họa cách sử dụng các thuật toán phân loại ở trên.