Đại học Quốc gia TPHCM
Trường Đại học Công nghệ Thông tin

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KỲ

CS116.M11.KHTN

NAVIE BAYES CLASSIFIER

Sinh viên thực hiện:

19520166 Phan Nhật Minh
19521287 Nguyễn Văn Chính
19520218 Nguyễn Minh Phú

Giáo viên hướng dẫn

TS. NGUYỄN VINH TIỆP

TPHCM, Tháng 12 năm 2021

Contents

1 Tổng quan **2**

2 Naive Bayes Classifier **2**

2.1 Định lý Bayes 2

2.2 Naive Bayes Classifier 2

2.3 Các phân phối thường dùng cho $p(x_i|C)$ 3

2.3.1 Gaussian Naive Bayes: 3

2.3.2 Multinomial Naive Bayes 3

2.3.3 Bernoulli Naive Bayes 4

3 Bài toán minh hoạ **4**

4 Ưu điểm và nhược điểm **5**

5 Ứng dụng **5**

1 Tổng quan

Classfication là 1 trong những bài toán rất phổ biến được sử dụng trong lĩnh vực máy học, từ đó mang lại rất nhiều đóng góp trong quá trình phát triển các ứng dụng cũng như nghiên cứu thuộc lĩnh vực này. Trong bài báo cáo này, chúng tôi xin giới thiệu về thuật toán Naive Bayes Classifier – một trong những thuật toán phân loại đơn giản và hiệu quả nhất giúp đưa ra dự đoán nhanh chóng.

2 Naive Bayes Classifier

Thuật toán Naive Bayes Classifier thuộc nhóm thuật toán học có giám sát. Về ý tưởng, thuật toán được xây dựng dựa trên cơ sở của định lý Bayes và hoạt động như một bộ phân loại theo xác suất, dự đoán trên cơ sở xác suất của một đối tượng. Để hiểu rõ hơn về thuật toán, đầu tiên chúng ta sẽ đến với khái niệm của định lý Bayes.

2.1 Định lý Bayes

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là "xác suất của A nếu có B". Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Công thức định lý Bayes: $P(A|B) = \frac{P(B|A).P(A)}{P(B)}$

2.2 Naive Bayes Classifier

Dựa trên định lý Bayes, thuật toán Naive Bayes Classifier xây dựng ý tưởng theo mô hình xác suất, từ đó sẽ đưa ra dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu.Trong Naive Bayes Classifier, ta cần giả định các yếu tố đầu vào là độc lập nhau

Để nắm rõ hơn cách hoạt động của thuật toán, ta sẽ xét ví dụ sau:

Xét bài toán classification với n classes $C_1, C_2, \dots C_n$. Giả sử có một điểm dữ liệu $x \in R^d$. Xác định x thuộc class nào.

- **Input**: tập các class $C_1, C_2, , C_n$ và điểm dữ liệu $x_1, x_2 \dots x_i$ cần phân loại
- **Output**: $\operatorname{argmax} p(C_i \mid x)$(với $i = 1, 2 \dots, n$) cho biết class i có $p(C_i \mid x)$ lớn nhất và đây chính là class cần tìm.

  Ta có:

$$\begin{aligned} \operatorname{argmax} p(C_i \mid x) &= \operatorname{argmax} \frac{p(x \mid C) \cdot p(c)}{p(x)} \\ &= \operatorname{argmax} p(x \mid C) \cdot p(c) \end{aligned}$$

Trong đó:

$$p(x|C) = p(x_1, x_2, ..., x_d|C) = \prod_{i=1}^{d} p(x_i|C)$$

- Các phân phối $p(C)$ và $p(x_i|C), i = 1, 2 \dots d$ sẽ được khởi tạo ban đầu và được cập nhật trong quá trình training
- Cách tính toán $p(x_i|C)$ sẽ phụ thuộc vào đặc điểm dữ liệu.

2.3 Các phân phối thường dùng cho $p(x_i|C)$

2.3.1 Gaussian Naive Bayes:

Sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

$$P(x_i = C) = P(x_i|\mu_{ci}, \sigma_{ci}^2) = \frac{1}{\sqrt{2\pi\sigma_{ci}^2}} \exp(-\frac{(x_i - \mu_{ci})^2}{2\sigma_{ci}^2})$$

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

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

Đây là cách tính của thư viện sklearn. Chúng ta cũng có thể đánh giá các tham số bằng MAP nếu biết trước priors của μ_{ci} và σ_{ci}^2

2.3.2 Multinomial Naive Bayes

Sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bag of Words.

Mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển.

Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó.

Khi đó $P(x_i|C)$ là tần suất từ thứ i xuất hiện trong các văn bản của class C.

$$\lambda_{ci} = P(x_i|C) = N_{\frac{N_{ci}}{N_C}}$$

Trong đó:

- N_{ci} là tổng số lần thứ i xuất hiện trong các văn bản của class C
- N_C là tổng số từ (kể cả lặp) xuất hiện trong class C

Để tránh trường hợp có một từ mới trong class C dẫn đến biểu thức (4) bằng 0, ta sử dũng kỹ thuật Laplace smoothing:

$$\lambda_{ci} = \frac{N_{ci} + \alpha}{N_C + d\alpha}$$

với $\alpha > 0$ và d là tổng số từ phân biệt.

2.3.3 Bernoulli Naive Bayes

Áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary – bằng 0 hoặc 1

$$P(x_i) = P(i|C)^{x_i} \cdot (1 - P(i|C))^{1-x_i}$$

3 Bài toán minh hoạ

Ta có 2 class A, B và tập train gồm một số câu được gán nhãn sẵn thuộc class nào. Dựa vào đó, hãy xác định câu test thuộc class nào

Set	Sentence	Content	Class
Training	s1	cookie milk candy cookie	A
	s2	cookie chocolate milk coffee	A
	s3	milk tea coffee	A
	s4	yogurt snack cake milk	B
Test	s5	cookie cookie chocolate snack	?

Trong tập train, ta có:
$P(A) = \frac{3}{4}$ và $P(B) = \frac{1}{4}$;
$d = 9$, là số từ phân biệt trong tập train

Dựa trên tập train, ta sẽ cập nhật xác suất của từng từ xuất hiện trong từng class. Cụ thể với từng class C ta sẽ tính $P(x_i|C)$ cho từng từ i (phân biệt) xuất hiện trong class đó theo công thức:

$$\frac{(N_i + 1)}{(N_C + d)}$$

với N_i là tổng số lần từ i xuất hiện trong các câu của class C, N_c là tổng số từ (kể cả lặp) xuất hiện trong class C.

Class A:

	cookie	milk	candy	chocolate	coffee	tea	Saigon	snack	cake
s_1	2	1	1	0	0	0	0	0	0
s_2	1	1	0	1	1	0	0	0	0
s_3	0	1	0	0	1	1	0	0	0
Total	3	3	1	1	2	1	0	0	0
$P(x_i \mid A)$	4/20	4/20	2/20	2/20	3/20	2/20	1/20	1/20	1/20

Class B

	cookie	milk	candy	chocolate	coffee	tea	yogurt	snack	cake
s_4	0	1	0	0	0	0	1	1	1
$P(x_i \mid B)$	1/13	2/13	1/13	1/13	1/13	1/13	2/13	2/13	2/13

Phần test, ta có:

- $P(s_{test} \mid A) = P(A).P(A \mid s_{test}) = \frac{3}{4}(\frac{4}{20})^2\frac{2}{20}\frac{1}{20} \approx 1,5.10^{-4}$
- $P(s_{test} \mid B) = P(B).P(B \mid s_{test}) = \frac{1}{4}(\frac{1}{13})^2\frac{1}{13}\frac{2}{13} \approx 1,75.10^{-5}$

$\Rightarrow P(s_{test}|A) > P(s_{test}|B)$

$\Rightarrow s_{test}$ thuộc class A

4 Ưu điểm và nhược điểm

Ưu điểm:

- NBC có thời gian training và test rất nhanh do giả định về tính độc lập giữa các thành phần
- Nếu giả sử về tính độc lập được thoả mãn, NBC được cho là cho kết quả tốt hơn so với SVM và logistic regression khi có ít dữ liệu training.
- NBC có thể hoạt động với các feature vector mà một phần là liên tục (sử dụng Gaussian Naive Bayes), phần còn lại ở dạng rời rạc (sử dụng Multinomial hoặc Bernoulli).
- Nó hoạt động tốt với dữ liệu nhiều chiều như phân loại văn bản, phát hiện thư rác email.

Nhược điểm:

- Giả định rằng tất cả các tính năng là độc lập không thường xảy ra trong cuộc sống thực vì vậy nó làm cho thuật toán bayes ngây thơ kém chính xác hơn các thuật toán phức tạp.

5 Ứng dụng

Thuật toán Naive Bayes Classifier được áp dụng vào một số loại ứng dụng sau:

- Real time Prediction: NBC chạy khá nhanh nên nó thích hợp áp dụng ứng dụng nhiều vào các ứng dụng chạy thời gian thực, như hệ thống cảnh báo, các hệ thống trading ...
- Multi class Prediction: Nhờ vào định lý Bayes mở rộng ta có thể ứng dụng vào các loại ứng dụng đa dự đoán, tức là ứng dụng có thể dự đoán nhiều giả thuyết mục tiêu.
- Text classification/ Spam Filtering/ Sentiment Analysis: NBC cũng rất thích hợp cho các hệ thống phân loại văn bản hay ngôn ngữ tự nhiên vì tính chính xác của nó lớn hơn các thuật toán khác.
- Recommendation System: Naive Bayes Classifier và Collaborative Filtering được sử dụng rất nhiều để xây dựng cả hệ thống gợi ý, ví dụ như xuất hiện các quảng cáo mà người dùng đang quan tâm nhiều nhất từ việc học hỏi thói quen sử dụng internet của người dùng...

References

[1] Machine Learning Co Ban - Tiep Huu Vu.
Machine Learning Co Ban Ebook: https://github.com/tiepvupsu/ebookMLCB

[2] Gaussian Naive Bayes by iq.opengenus.org
https://iq.opengenus.org/gaussian-naive-bayes/

[3] Bernoulli Naive Bayes by iq.opengenus.org
https://iq.opengenus.org/bernoulli-naive-bayes/