Naïve Bayes classifier

Taehoon Ko (thoon.koh@gmail.com)

목표

- **다음을 이해한다.**
  - Bayes theorem과 사후확률 (posterior probability)
  - “Naïve”한 방법으로 베이즈 분류기를 만드는 이유
  - 나이브 베이즈 분류기 (Naïve Bayes classifier)

Bayes theorem

- 조건부 확률 (Conditional Probability)

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \qquad P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

- 베이즈 정리 (Bayes theorem)

likelihood of the evidence 'X' if 'Y' is given

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

prior probability of 'Y'

posterior probability of 'Y' given the evidence 'X'

prior probability that the evidence 'X' itself is given.

Bayes theorem

- 조건부 확률 (Conditional Probability)

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \qquad P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

- 베이즈 정리 (Bayes theorem)

가능도
: (결과 Y가 관측되었을 때) 현상 X가 나타날 가능성

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

사전확률
: (X가 관측되기 전에 파악한) 결과 Y의 확률

사후확률
: 현상 X가 주어졌을 때, 결과 Y의 확률

증거
: 현상 X가 나타날 확률

Example of Bayes theorem

- Given:
  - A doctor knows that **meningitis(뇌막염)** causes **stiff neck(류머티즘)** 50% of the time ➔ Likelihood
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Exact Bayes classifier

- A probabilistic framework for solving classification problems
  - Consider each attribute and class label as random variables
    - The goal is to predict class of given new point $(X_1, X_2,\ldots, X_p)$
    - Specifically, we want to find the value of Y that maximizes $P(C \mid X_1, X_2,\ldots, X_p)$

$$\rightarrow C = \underset{C_j}{\operatorname{argmax}}\, P\left(C_j \mid X_1,\ X_2,\ldots,\ X_p\right)$$

- Problem: How to estimate $\boldsymbol{P(C \mid X_1, X_2,\ldots, X_p)}$ directly from data?

Exact Bayes classifier

- How to estimate $\boldsymbol{P(C \mid X_1, X_2,\ldots, X_p)}$
  - Compute the posterior probability $P(C \mid X_1, X_2,\ldots, X_p)$ for all values of C using the Bayes theorem.

$$P(C \mid X_1, X_2, \ldots, X_p) = \frac{P(X_1, X_2, \ldots, X_p \mid C)P(C)}{P(X_1, X_2, \ldots, X_p)}$$

  - Suppose that there are 2 classes C_1, C_2.
  - In order to predict a class of given new record $(X_1, X_2,\ldots, X_p)$, following two probabilities are compared.

$$P(C_1 \mid X_1, X_2, \ldots, X_p) \quad \text{vs.} \quad P(C_2 \mid X_1, X_2, \ldots, X_p)$$

Exact Bayes classifier

- How to estimate $\boldsymbol{P(C \mid X_1, X_2,\ldots, X_p)}$

$$P(C_1 \mid X_1, X_2, \ldots, X_p) = \frac{P(X_1, X_2, \ldots, X_p \mid C_1)P(C_1)}{P(X_1, X_2, \ldots, X_p)}$$

$$P(C_2 \mid X_1, X_2, \ldots, X_p) = \frac{P(X_1, X_2, \ldots, X_p \mid C_2)P(C_2)}{P(X_1, X_2, \ldots, X_p)}$$

  - Both probabilities include a evidence term $P(X_1, X_2,\ldots, X_p)$.
  - Choosing value of C that maximizes $P(C \mid X_1, X_2,\ldots, X_p)$ is equivalent to choosing value of C that maximizes $P(X_1, X_2,\ldots, X_p|C)P(C)$.

- Problem: How to estimate $\boldsymbol{P(X_1, X_2,\ldots, X_p|C)}$ and $\boldsymbol{P(C)}$?

How to estimate $P(X_1, X_2,..., X_p|C)$

- Unfortunately, you cannot always estimate $P(X_1, X_2,..., X_p|C)$.
  - If input variables are binary, data should contain 2^p combinations of input values. ➔ Unrealistic
  - Example:

Rain	Temperature	Humidity	Play?
No	Low	Low	Yes
No	High	Mid	No
Yes	Mid	Mid	No
Yes	High	Low	Yes
Yes	Low	High	No
No	Mid	High	Yes
No	High	High	No

P(X1='No', X2='Low', X3='Low' | Y='Yes') = 1/3

P(X1='Yes', X2='Low', X3='Low' | Y='Yes') = Not available

How to estimate $P(X_1, X_2,..., X_p|C)$

- Assume independence among input variables X_js when class is given:
  - $P(X_1, X_2,..., X_p|C_i) = P(X_1|C_i) \times P(X_2|C_i) \times \cdots \times P(X_p|C_i) = \prod_{j=1}^{p} P(X_j|C_i)$
  - $P(X_j|C_i)$ for all X_j and C_i can be estimated using training set.

> We can estimate $\boldsymbol{P(X_1, X_2,..., X_p|C)}$ approximately by assuming independence among input variables $\boldsymbol{X_j}$s.

- With this assumption, exact Bayes classifier is changed to "naïve Bayes classifier"

How to estimate $P(C)$

- We can estimate P(C) using the class distribution of training set.
  - Example
    - 2 classes C_1, C_2.
    - Number of points in training set = 1,000
    - Number of points with the class C_1 = 400
    - Number of points with the class C_2 = 600
  - $P(C_1) = \frac{400}{1000} = 0.4$
  - $P(C_2) = \frac{600}{1000} = 0.6$

Naïve Bayes classifier

- Assumption
  - Independences among input variables
- How to train naïve Bayes classifier
  - Given training set, calculate
    - $P(X_j|C_i)$ for all input variables X_j and classes C_i
    - $P(C_i) = \frac{\text{number of points with class } C_i}{\text{number of training points}}$ for all classes C_i
- How to predict a new point

$$\text{predicted class} = \arg\max_{C_i} P(C_i \mid X_1, X_2, \ldots, X_p)$$

$$= \arg\max_{C_i} P(C_i) \prod_{j=1}^{p} P(X_j \mid C_i)$$

How to estimate probabilities from data?

ID	Refund	Marital Status	Taxable Income ($)	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(C_i) = |C_i|/N$
  - where $|C_i|$ is a number of points belonging to class C_i
  - ex) P(No) = 7/10, P(Yes) = 3/10
- For discrete attributes:

  $P(X_j|C_i) = |X_{ji}|/|C_i|$

  - where $|X_{ji}|$ is number of points having input variable X_j and belonging to class C_i
  - Examples:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

How to estimate probabilities from data?

- For continuous input variables:
  - Discretize the range into bins
    - one ordinal input variable per bin
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new input variable
  - Probability density estimation:
    - Assume that input variable follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, we can use it to estimate the conditional probability $P(X_j|C_i)$

How to estimate probabilities from data?

- Normal distribution (or Gaussian distribution)

ID	Refund	Marital Status	Taxable Income ($)	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(X_j \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}^2}} e^{-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}}$$

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> predicted_class = No

Variations of naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

Original: $P(X_i \mid C) = \frac{N_{ic}}{N_c}$

Laplace: $P(X_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$

m-estimate: $P(X_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$

c: number of classes

p: prior probability

m: parameter

Example of naïve Bayes classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

정리

- **Naïve Bayes classifier의 특징**
  - 노이즈에 민감하지 않음 (노이즈에 강건함 = Robust to noises)
  - 출력 변수와 큰 연관성이 없는 입력 변수에 강건함
  - (확률분포 추정에 크게 지장을 주지 않는 수준에서) 데이터에 결측치 (missing value) 가 있어도 학습이 가능함
  - 문서-단어 행렬 (Document-term matrix) 과 같은 희소행렬 (sparse matrix) 에 강하다고 알려짐 ➔ 문서 분류 모델 학습에 용이하다는 의견 많음
  - 머신러닝에서 가장 기본적인 generative model

정리

- Naïve Bayes classifier의 특징
  - 입력 변수의 독립성 가정이 성립하지 않는 데이터에 매우 취약함
    - 그렇다면 (exact) Bayes classifier를 사용?
    - Bayesian belief network (BBN) 과 같은 방법 고려
  - (경험적으로) 다른 머신러닝 알고리즘에 의해 학습한 모델보다 성능이 떨어짐
  - Naïve Bayes classifier에서 계산하는 사후 확률 (posterior probability) 은 절대로 정확한 확률이 아님!
    - 독립 가정으로 여러 확률 값을 곱하기 때문에 실제 사후 확률보다 상당히 낮게 계산되는 경향