Naïve Bayes classifier

고태훈 (taehoonko@dm.snu.ac.kr)

Bayes theorem

- **Conditional Probability:**

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- **Bayes theorem:**

likelihood of the evidence 'A' if C is given

prior probability of C

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

posterior probability of 'C' given the evidence 'A'

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

Example of Bayes theorem

- **Given:**
  - A doctor knows that **meningitis(뇌막염)** causes **stiff neck(류머티즘)** 50% of the time
  - 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 $P(C \mid X_1, X_2, \ldots, X_p)$

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

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

Exact Bayes classifier

❖ How to estimate $P(C \mid X_1, X_2,\ldots, X_p)$

- 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 $P(C \mid X_1, X_2, \ldots, X_p)$**

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

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

  - 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 $P(X_1, X_2, \ldots, X_p \mid C) P(C)$.

- **Problem: How to estimate $P(X_1, X_2, \ldots, X_p \mid C)$ and $P(C)$?**

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

- **Unfortunately, you cannot always estimate** $P(X_1, X_2, \ldots, X_p | C)$.
  - 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 $P(X_1, X_2, \ldots, X_p|C)$

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

We can estimate $P(X_1, X_2, \ldots, X_p|C)$ approximately by assuming independence among input variables X_js.

- 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?

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

- **Normal distribution:**

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

- **For (Income, Class=No):**
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

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

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) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 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:

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

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

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

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

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

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

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

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

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

Naïve Bayes (Summary)

Pros

- Robust to isolated noise points
- Robust to irrelevant input variables
- Able to handle missing values by ignoring the points containing missing values during probability estimate calculations
- Good performance on sparse data such as document-term matrix

Cons

- Independence assumption may not hold for some input variables
  - Use other techniques such as Bayesian Belief Networks (BBN)
- If the purpose is to be actually estimated the probability belong to each class, naïve Bayes leads to a very biased results.
- Need sufficient number of data points