Mahidol University
Faculty of Medicine Ramathibodi Hospital
Section for Clinical Epidemiology and Biostatistics

Naïve Bayes Classifier

Ratchainant Thammasudjarit, Ph.D.

Bayes Theorem

Priors, Likelihood, Marginal, and Posterior

- Let A and B be random variables. Bayes' theorem is defined as follows

Likelihood (we can observe)

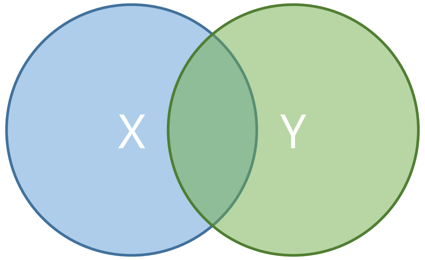
Posterior (we want to estimate) $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Prior (we know it)

Marginal (we can either directly observe or do some math)

Alternative Form of Bayes' Theorem

Dealing with marginal

- From the probability theory



$$P(X) = P(X \cap Y) + P(X \cap Y')$$

$$= P(X, Y) + P(X, \sim Y) \qquad \text{Eq.1}$$

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \qquad \text{Eq.2}$$

Hence

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

Applications of Bayes' Theorem to Healthcare

Liver disease

- Scenario

  Historical data tells that 10% of patients visiting our clinic have liver disease. 7% of patients diagnosed with liver disease, they are alcoholics. 5% of patients according to the test are alcoholics. Finding out the probability of liver disease if a given patient is alcoholics.

Prior: Historical data tells that 10% of patients visiting our clinic have liver disease, P(liver disease) = 0.1

Likelihood: 7% of patients diagnosed with liver disease, they are alcoholics, P(alcoholics | liver disease) = 0.07

Marginal: 5% of patients according to the test are alcoholics, P(alcoholics) = 0.05

Posterior: P(liver disease | alcoholics) = $(0.07 \times 0.1)/0.05 = 0.14$

Conditional Independence

Assumption of Naïve Bayes

- Recall probability theory
  - Let A and B be random variables
  - Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Or eventually

$$P(A|B) = P(A)$$

Conditional Independence

Assumption of Naïve Bayes

- Definition: Two events A and B are conditionally independent given an event C with $P(C) > 0$ if

$$P(A \cap B|C) = P(A|C)P(B|C)$$

- Recall that from the definition of conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Independence

Assumption of Naïve Bayes

- if $P(B) > 0$, by conditioning on C, we obtain

$$P(A|B,C) = \frac{P(A \cap (B \cap C))}{P(B \cap C)}$$

$$= \frac{P(A \cap B \cap C)}{P(B \cap C)} \cdot \frac{P(C)}{P(C)}$$

$$= \frac{P(A \cap B|C)}{P(B|C)}$$

Conditional Independence

Assumption of Naïve Bayes

- if $P(B|C)$ and $P(C) \neq 0$ and if A and B are conditionally independent given C, we obtain

$$P(A|B,C) = \frac{P(A \cap B|C)}{P(B|C)}$$

$$= \frac{P(A|C)P(B|C)}{P(B|C)}$$

$$= P(A|C)$$

Translation:

- Knowing prior B does not improve posterior of A given C

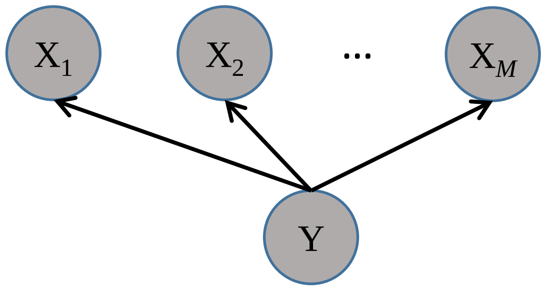
Example: Lung Cancer prediction from smoking and sex

- Knowing prior probability of sex does not improve posterior probability of lung cancer given smoking

Naïve Bayes Model

Model Representation

- Two choices
  - Directed Graph
  - Plate Notations



Directed Graph

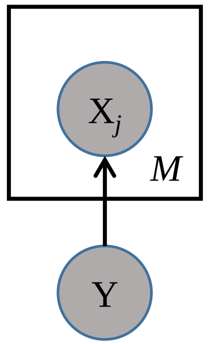


Plate Graph

Naïve Bayes Model

Equation

- Let Y be a set of k class labels, $Y = \{y_1, y_2 \ldots y_k\}$ and x_j be any possible value of $X_j, 1 \leq j \leq M$
- For each class label i that $1 \leq i \leq k$, Naïve Bayes Model is defined as follows

$$P(y_i|x_1, x_2, \ldots, x_M) = \frac{P(x_1, x_2, \ldots, x_M|y_i)P(y_i)}{P(x_1, x_2, \ldots, x_M)}$$

$$\propto P(x_1, x_2, \ldots, x_M|y_i)P(y_i)$$

Assuming X_1, X_2, … , X_M are conditionally independence given Y

$$P(x_1, x_2, \ldots, x_M|y_i) = P(x_1|y_i)P(x_2|y_i) \ldots P(x_M|y_i)$$

$$P(y_i|x, x_2, \ldots, x_M) \propto P(y_i)\prod_{j=1}^{M} P(x_j|y_i)$$

Naïve Bayes Model

MLE Inference

- From training data $N \times M$, the maximum likelihood estimator for Naïve Bayes model is defined as follows

$$\hat{y}_i = \arg\max_{y_i \in Y} \prod_{j=1}^{M} P(x_j|y_i)$$

Apply logarithmic function to avoid overflow problem

$$\hat{y}_i \approx \arg\max_{y_i \in Y} \log \prod_{j=1}^{M} P(x_j|y_i)$$

$$\approx \arg\max_{y_i \in Y} \sum_{j=1}^{M} \log P(x_j|y_i)$$

Or a convenience form

$$\hat{y} \approx \arg\max_{y \in Y} \sum_{j} \log P(x_j|y)$$

Naïve Bayes Model

MAP Inference

- From training data $N \times M$, the Maximum A-Posteriori estimator for Naïve Bayes model is defined as follows

$$\hat{y}_i = \arg\max_{y_i \in Y} \prod_{j=1}^{M} P(y_i|x_j)$$

$$\approx \arg\max_{y_i \in Y} \prod_{j=1}^{M} P(x_j|y_i)P(y_i)$$

$$\approx \arg\max_{y_i \in Y} P(y_i) \prod_{j=1}^{M} P(x_j|y_i)$$

$$\approx \arg\max_{y_i \in Y} \log\left(P(y_i) \prod_{j=1}^{M} P(x_j|y_i) \right)$$

MLE VS MAP

Prior

- Use MAP when prior is taking into account

$$\hat{y}_i \approx \arg\max_{y \in Y} \sum_j \log P(x_j|y)$$

MLE

$$\hat{y}_i \approx \arg\max_{y_i \in Y} \log\left(P(y_i) \prod_{j=1}^{M} P(x_j|y_i) \right)$$

MAP

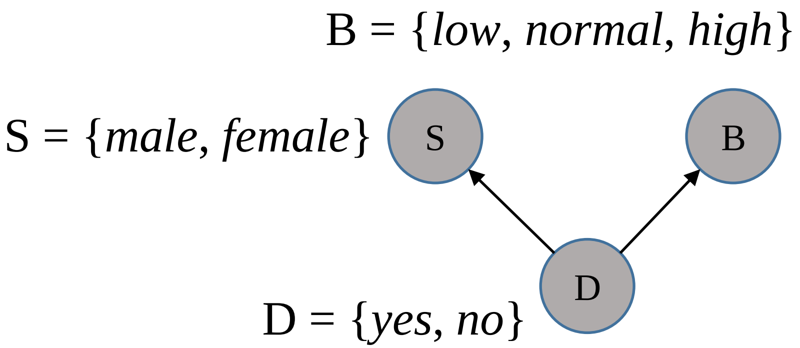
Example

Diabetes Prediction

Blood Pressure (B)	**Sex (S)**	**Diabetes (D)**
normal	*male*	*no*
low	*female*	*no*
high	*male*	*yes*
normal	*female*	*yes*
normal	*male*	*no*
low	*female*	*no*
high	*male*	*yes*

- From training data, the model is designed as follows

- Model parameters are
- $P(S \mid D)$
- $P(B \mid D)$
- $P(D)$



Example

Learning

- Step 1: Create frequency tables
  - #S
  - #S,D
  - #B,D

D	**#**
yes	3
no	4

S	**D**	**#**
male	*yes*	2
male	*no*	2
female	*yes*	1
female	*no*	2

B	**D**	**#**
low	*yes*	0
low	*no*	2
normal	*yes*	1
normal	*no*	2
high	*yes*	2
high	*no*	0

Example

Learning

- Step 2: Initialize the joint probability tables

D	*P*(D)
yes	3/7
no	4/7

S	D	*P*(S,D)
male	*yes*	2/7
male	*no*	2/7
female	*yes*	1/7
female	*no*	2/7

B	D	P(B,D)
low	*yes*	0/7
low	*no*	2/7
normal	*yes*	1/7
normal	*no*	2/7
high	*yes*	2/7
high	*no*	0/7

Zero probability is undesirable. Smoothing probability will be applied to the last table (B,D)

Example

Learning

- Step 2.1: Smoothing probability
  - $\alpha = 0.1$

D	*P*(D)
yes	3/7
no	4/7

S	D	*P*(S,D)
male	*yes*	2/7
male	*no*	2/7
female	*yes*	1/7
female	*no*	2/7

B	D	P_{smooth}(B,D)
low	*yes*	0.0132
low	*no*	0.2763
normal	*yes*	0.1447
normal	*no*	0.2763
high	*yes*	0.2763
high	*no*	0.0132

Example

Learning

- Step 3: Calculate the conditional probability table
  - $P(S|D)$

D	P(D)
yes	3/7
no	4/7

S	D	P(S,D)
male	*yes*	2/7
male	*no*	2/7
female	*yes*	1/7
female	*no*	2/7

S	D	P(S|D)
male	*yes*	(2/7)÷(3/7) = 2/3
male	*no*	(2/7)÷(4/7) = 1/2
female	*yes*	(1/7)÷(3/7) = 1/3
female	*no*	(2/7)÷(4/7) = 1/2

Example

Learning

- Step 3: Calculate the conditional probability table
  - $P(B|D)$

D	*P*(D)
yes	3/7
no	4/7

B	D	P_{smooth}(B,D)
low	*yes*	0.0132
low	*no*	0.2763
normal	*yes*	0.1447
normal	*no*	0.2763
high	*yes*	0.2763
high	*no*	0.0132

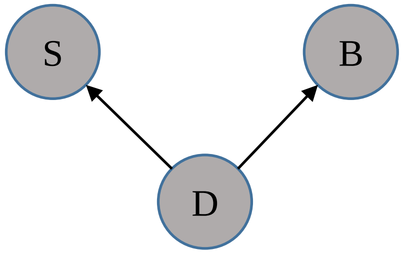
B	D	P(B|D)
low	*yes*	0.0132 ÷ (3/7) = 0.6447
low	*no*	0.2763 ÷ (4/7) = 0.2533
normal	*yes*	0.1447 ÷ (3/7) = 0.0307
normal	*no*	0.2763 ÷ (4/7) = 0.4836
high	*yes*	0.2763 ÷ (3/7) = 0.6447
high	*no*	0.0132 ÷ (4/7) = 0.0230

Example

Learning

- Your model

S	D	*P*(S|D)
male	*yes*	0.6666
male	*no*	0.5000
female	*yes*	0.3333
female	*no*	0.5000



B	D	P(B|D)
low	*yes*	0.6447
low	*no*	0.2533
normal	*yes*	0.0307
normal	*no*	0.4836
high	*yes*	0.6447
high	*no*	0.0230

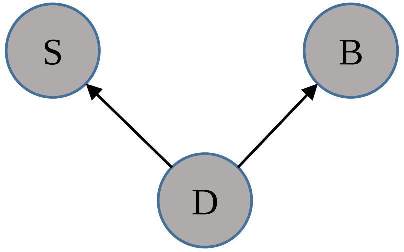
D	*P*(D)
yes	0.4286
no	0.5714

Note: In practice, for fast learning, model parameters are stored in the form of joint probability. Such parameters will be transform to conditional probability before inference.

Example

Inference

S	D	P(S|D)
male	*yes*	0.6666
male	*no*	0.5000
female	*yes*	0.3333
female	*no*	0.5000



B	D	P(B|D)
low	*yes*	0.6447
low	*no*	0.2533
normal	*yes*	0.0307
normal	*no*	0.4836
high	*yes*	0.6447
high	*no*	0.0230

D	P(D)
yes	0.4286
no	0.5714

- Given a patient is male and has high blood pressure, what is the conclusions for his diabetes?

$$\log P(male|yes) + \log P(high|yes) = \log 0.6666 + \log 0.6447 = -0.8445$$

$$\log P(male|no) + \log P(high|no) = \log 0.5000 + \log 0.0230 = -4.4654$$

With MLE inference, this patient is likely to have diabetes

Coding Practice