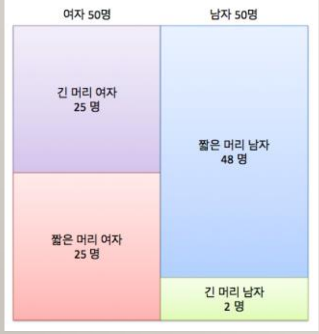
PNU Industrial Data Science

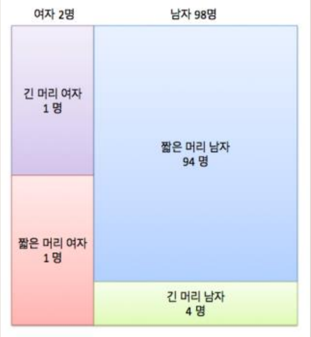
Naïve Bayes

Prof. Hyerim Bae

남자일까요? 여자일까요?

$p(man|long\ hair)$





Source: https://brunch.co.kr/@chris-song/59

Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

01 ~~Tree for Classification~~

02 ~~Measure for purity~~

03 ~~Decision Tree~~

Naïve Bayes: The Basic Idea

For a given new record to be classified, find other records like it (i.e., same values for the predictors)

What is the prevalent class among those records?

Assign that class to your new record

Characteristics

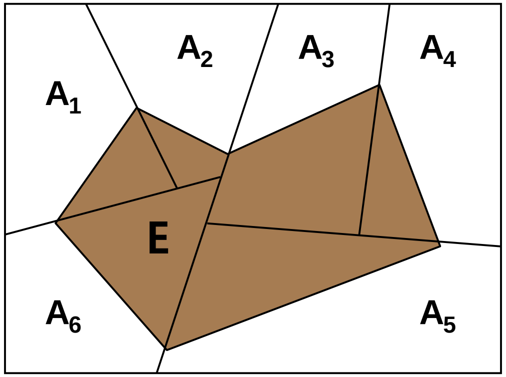
Data-driven, not model-driven

Make no assumptions about the data

Bayes Rule

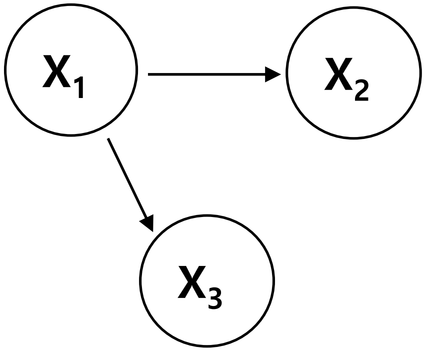
$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

$$p(A_i|E) = \frac{p(E|A_i)p(A_i)}{P(E)} = \frac{p(E|A_i)p(A_i)}{\sum_i p(E|A_i)p(A_i)}$$

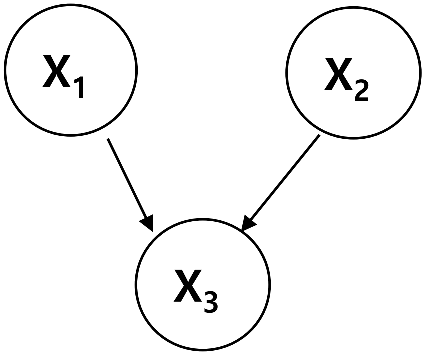


- Based on definition of conditional probability
- $p(A_i|E)$ is posterior probability given evidence E
- $p(A_i)$ is the prior probability
- $P(E|A_i)$ is the likelihood of the evidence given A_i
- p(E) is the preposterior probability of the evidence

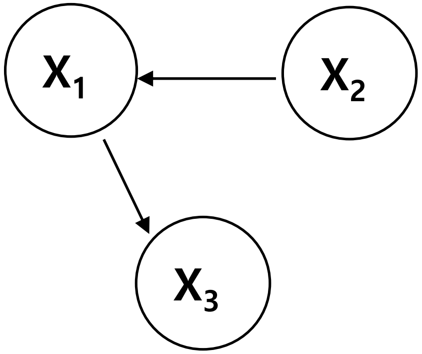
Arc Reversal - Bayes Rule



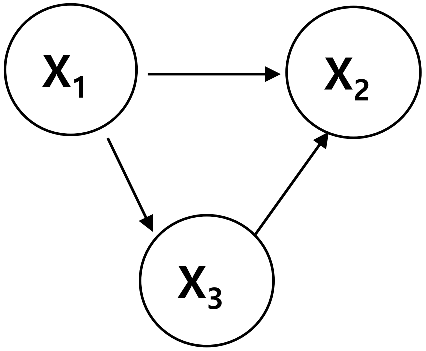
$$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$$



$$p(x_1, x_2, x_3) = p(x_3 \mid x_2, x_1)\, p(x_2)\, p(x_1)$$



$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_3 \mid x_1)\, p(x_2 , x_1) \\ &= p(x_3 \mid x_1)\, p(x_1 \mid x_2)\, p(x_2) \end{aligned}$$



$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_3, x_2 \mid x_1)\, p(x_1) \\ &= p(x_2 \mid x_3, x_1)\, p(x_3 \mid x_1)\, p(x_1) \end{aligned}$$

Cutoff Probability Method

- Establish cutoff probability for the class of interest
  - 어떤 레코드가 특정 클래스에 속한다고 판단하는 기준값이 될 확률값을 설정
- Find all the training records just like the new record
  - 새로운 레코드와 비슷한 모든 학습 레코들들을 찾는다.
- Determine the probability that those records belong to the class of interest
  - 그 레코들들이 관련 클래스에 속할 확률을 결정
- If the probability is above the cutoff probability, assign the new record to the class of interest
  - 만약 확률이 기준값을 넘으면 해당 클래스에 속한다고 결정

Conditional Probability

- The probability event A occurs, when event B occurs:

$$P(A|B)$$

  - The probability of being a class when predictors have certain values
    Predictors 가 특정한 값을 가질 때, 클래스에 속할 확률

$$P(C_i|x_1, \cdots, x_p)$$

  - Choose *i* with the biggest probability
    '가장 큰 확률을 가지는 클래스에 소속된다'라고 판정!
    베이지안 분류기는 범주형 예측기들에만 작동

Usage

- Requires categorical variables

- Numerical variable must be binned and converted to categorical

- Can be used with very large data sets

- Example: Spell check programs assign your misspelled word to an established “class” (i.e., correctly spelled word)

Example

	Prior Legal (x_1)	No Prior Legal (x_2)	Total
Fraudulent (C_1)	50	50	100
Truthful (C_2)	180	720	900
Total	230	770	1,000

- The probability of fraudulent if it has an prior experience of (새로운 회사가 이전에 법적 문제가 있었다면, 부정 클래스에 속할 확률) = 50/230

- Using the “Assign to the Most Probable Class” Method
  - 모든 레코드를 정직으로 분류하게 됨
- Using the Cuttoff Probability Method
  - Cutoff 가 0.20인 경우
  - 50/230 = 0.22 => 법적문제를 야기한 경우 부정한 것으로 판정

$$P(C_i|x_1, x_2, \cdots, x_p) = \frac{P(x_1, x_2, \cdots, x_p|C_i)P(C_i)}{P(x_1, x_2, \cdots, x_p|C_1)P(C_1) + \cdots + P(x_1, x_2, \cdots, x_p|C_m)P(C_m)}$$

Exact Bayes Classifier

Relies on finding other records that share same predictor values as record-to-be-classified.

Want to find “probability of belonging to class *C*, given specified values of predictors.”

Even with large data sets, may be hard to find other records that **exactly match** your record, in terms of predictor values.

미국 중서부의 고수익을 갖는 히스패닉 남성으로 지난선거에서 투표했지만, 그 이전 선거에서는 투표하지 않았고, 세 명의 딸과 한 명의 아들이 있으면 현재, 이혼한 사람

Solution – Naïve Bayes

- Assume independence of predictor variables (within each class)

- Use multiplication rule

- Find same probability that record belongs to class C, given predictor values, without limiting calculation to records that share all those same values

Calculations

1. For a given new record to be classified, find other records like it (i.e., same values for the predictors)
2. What is the prevalent class among those records?
3. Assign that class to your new record

1. For class1, find the individual probabilities that each predictor value occurs in class 1(클래스 1에 대해서 각 예측변수 값이 그 클래스에서 발생할 확률을 구한다.
2. Multiply these probability times each other, then times the proportion of records belonging to class 1 (이러한 확률값들을 곱한다.)
3. Repeat steps 1 and 2 for all the classes
4. Estimate a probability for class *i* by taking the value calculated in step 2 for class *i* and dividing it by the sum of such values for all classes
5. Assign the record to the class with the highest probability value for this set of predictor values

$$P_{nb}(C_1|, \cdots, x_p)$$

$$=\frac{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)]}{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)]+\cdots+P(C_m)\ [P(x_1|C_m)P(x_2|C_m)\cdots P(x_p|C_m)]}$$

Problem of Formidable(복잡한) formula

- Need to simplify
  - The probability of a class with given predictors
  - If we are interested only in the ranking…

Multiplication of probabilities of a record with a predictor in each class

Probability of a class

$$P_{nb}(C_1|,\cdots,x_p)$$

$$=\frac{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)]}{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)]+\cdots+P(C_m)\ [P(x_1|C_m)P(x_2|C_m)\cdots P(x_p|C_m)]}$$

The probability of predictors in every class

Example: Financial Fraud

Target variable: Audit finds fraud, no fraud

Predictors:

- Prior pending legal charges (yes/no)
- Size of firm (small/large)

Charges?	Size	Outcome
y	small	truthful
n	small	truthful
n	large	truthful
n	large	truthful
n	small	truthful
n	small	truthful
y	small	fraud
y	large	fraud
n	large	fraud
y	large	fraud

Exact Bayes Calculations

Goal: classify (as “fraudulent” or as “truthful”) a small firm with charges filed

There are 2 firms like that, one fraudulent and the other truthful

P(fraud | charges=y, size=small) = ½ = 0.50

Note: calculation is limited to the two firms matching those characteristics

Naïve Bayes Calculations

Same goal as before

Compute 2 quantities:

- Proportion of “charges = y” among frauds, times proportion of “small” among frauds, times proportion frauds = 3/4 * 1/4 * 4/10 = 0.075
- Prop “charges = y” among frauds, times prop. “small” among truthfuls, times prop. truthfuls = 1/6 * 4/6 * 6/10 = 0.067

P(fraud | charges, small) = 0.075/(0.075+0.067)

= 0.53

Naïve Bayes, cont.

- Note that probability **estimate** does not differ greatly from **exact**
- All records are used in calculations, not just those matching predictor values
- This makes calculations practical in most circumstances
- Relies on assumption of independence between predictor variables within each class

Independence Assumption

- Not strictly justified (variables often correlated with one another)

- Often “good enough”

Advantages

- Handles purely categorical data well
- Works well with very large data sets
- Simple & computationally efficient

Shortcomings

- Requires large number of records

- Problematic when a predictor category is not present in training data
  - Assigns 0 probability of response, ignoring information in other variables