



Institut Informatika & Bisnis
DARMAJAYA
Yayasan Alfian Husin



**Kampus
Merdeka**
INDONESIA JAYA

**MERDEKA
BELAJAR**

DATA SCIENCE DARMAJAYA
“YOUR BEST FUTURE IN DATA”

MEETING: [11]

INTRODUCTION TO NAÏVE BAYES ALGORITHM

BY: HENDRA KURNIAWAN



Introduction to Naïve Bayes Algorithm

1. Bayesian Classification
2. Bayes Rule
3. Bayesian Classifier
4. Naïve Bayes Classifier
5. How to Estimate Probabilities from Data



Bayesian Classification: Why?

- A statistical classifier: performs *probabilistic prediction*, i.e., predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

Bayes' Rule

$$p(h | d) = \frac{P(d | h)P(h)}{P(d)}$$

Who is who in Bayes' rule

Understanding Bayes' rule

d = data

h = hypothesis (model)

- rearranging

$$p(h | d)P(d) = P(d | h)P(h)$$

$$P(d, h) = P(d, h)$$

the same joint probability

on both sides

$P(h)$: prior belief (probability of hypothesis h before seeing any data)

$P(d | h)$: likelihood (probability of the data if the hypothesis h is true)

$P(d) = \sum_h P(d | h)P(h)$: data evidence (marginal probability of the data)

$P(h | d)$: posterior (probability of hypothesis h after having seen the data d)

Example of Bayes Theorem

- Given:
 - A doctor knows that meningitis causes stiff neck 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?

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

Choosing Hypotheses

- *Maximum Likelihood* hypothesis:

$$h_{ML} = \arg \max_{h \in H} P(d | h)$$

- Generally we want the most probable hypothesis given training data. This is the *maximum a posteriori* hypothesis:
 - Useful observation: it does not depend on the denominator $P(d)$

$$h_{MAP} = \arg \max_{h \in H} P(h | d)$$



Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2, \dots, A_n)
 - Goal is to predict class C
 - Specifically, we want to find the value of C that maximizes $P(C | A_1, A_2, \dots, A_n)$
- Can we estimate $P(C | A_1, A_2, \dots, A_n)$ directly from data?

Bayesian Classifiers

- Approach:
 - compute the posterior probability $P(C | A_1, A_2, \dots, A_n)$ for all values of C using the Bayes theorem

$$P(C | A_1 A_2 \dots A_n) = \frac{P(A_1 A_2 \dots A_n | C) P(C)}{P(A_1 A_2 \dots A_n)}$$

- Choose value of C that maximizes $P(C | A_1, A_2, \dots, A_n)$
- Equivalent to choosing value of C that maximizes $P(A_1, A_2, \dots, A_n | C) P(C)$
- How to estimate $P(A_1, A_2, \dots, A_n | C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
 - $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C_j) P(A_2 | C_j) \dots P(A_n | C_j)$
 - Can estimate $P(A_i | C_j)$ for all A_i and C_j .
 - This is a simplifying assumption which may be violated in reality
- The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier

$$c_{Naive\ Bayes} = \arg \max_c P(c)P(\mathbf{x} | c) = \arg \max_c P(c) \prod_i P(a_i | c)$$

How to Estimate Probabilities from Data?

<i>Tid</i>	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: $P(C) = N_c / N$

- e.g., $P(\text{No}) = 7/10$,
 $P(\text{Yes}) = 3/10$

- For discrete attributes:

$$P(A_i | C_k) = |A_{ik}| / N_c$$

- where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k

- Examples:

$$P(\text{Status}=\text{Married} | \text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes} | \text{Yes})=0$$

How to Estimate Probabilities from Data?

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

How to Estimate Probabilities from Data?

Tid	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:

$$P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (A_i, c_j) pair
- For (Income, Class=No):
 - If Class=No
 - sample mean = 110
 - sample variance = 2975

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

Naïve Bayesian Classifier: Training Dataset

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

New Data:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



Naïve Bayesian Classifier: An Example

Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X | C_i)P(C_i)$, for $i=1,2$

First step: Compute $P(C)$ The prior probability of each class can be computed based on the training tuples:

$$P(\text{buys_computer=yes})=9/14=0.643$$

$$P(\text{buys_computer=no})=5/14=0.357$$



Naïve Bayesian Classifier: An Example

Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X | C_i)P(C_i)$, for $i=1,2$

Second step: compute $P(X | C_i)$

$$\begin{aligned} P(X | \text{buys_computer=yes}) &= P(\text{age=youth} | \text{buys_computer=yes}) \times \\ &\quad P(\text{income=medium} | \text{buys_computer=yes}) \times \\ &\quad P(\text{student=yes} | \text{buys_computer=yes}) \times \\ &\quad P(\text{credit_rating=fair} | \text{buys_computer=yes}) \\ &= 0.044 \end{aligned}$$

$$P(\text{age=youth} | \text{buys_computer=yes}) = 0.222$$

$$P(\text{income=medium} | \text{buys_computer=yes}) = 0.444$$

$$P(\text{student=yes} | \text{buys_computer=yes}) = 6/9 = 0.667$$

$$P(\text{credit_rating=fair} | \text{buys_computer=yes}) = 6/9 = 0.667$$



Naïve Bayesian Classifier: An Example

Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X | C_i)P(C_i)$, for $i=1,2$

Second step: compute $P(X | C_i)$

$$\begin{aligned} P(X | \text{buys_computer=no}) &= P(\text{age=youth} | \text{buys_computer=no}) \times \\ &\quad P(\text{income=medium} | \text{buys_computer=no}) \times \\ &\quad P(\text{student=yes} | \text{buys_computer=no}) \times \\ &\quad P(\text{credit_rating=fair} | \text{buys_computer=no}) \\ &= 0.019 \end{aligned}$$

$$P(\text{age=youth} | \text{buys_computer=no}) = 3/5 = 0.666$$

$$P(\text{income=medium} | \text{buys_computer=no}) = 2/5 = 0.400$$

$$P(\text{student=yes} | \text{buys_computer=no}) = 1/5 = 0.200$$

$$P(\text{credit_rating=fair} | \text{buys_computer=no}) = 2/5 = 0.400$$



Naïve Bayesian Classifier: An Example

Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X | C_i)P(C_i)$, for $i=1,2$

We have computed in the first and second steps:

$$P(\text{buys_computer}=\text{yes})=9/14=0.643$$

$$P(\text{buys_computer}=\text{no})=5/14=0.357$$

$$P(X | \text{buys_computer}=\text{yes})= 0.044$$

$$P(X | \text{buys_computer}=\text{no})= 0.019$$

Third step: compute $P(X | C_i)P(C_i)$ for each class

$$P(X | \text{buys_computer}=\text{yes})P(\text{buys_computer}=\text{yes})=0.044 \times 0.643=0.028$$

$$P(X | \text{buys_computer}=\text{no})P(\text{buys_computer}=\text{no})=0.019 \times 0.357=0.007$$

The naïve Bayesian Classifier predicts X belongs to class (“**buys_computer = yes**”)

Example

Training set :

<i>Tid</i>	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

Given a Test Record:

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

Given a Test Record:

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

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No: sample mean=110
 sample variance=2975

If class=Yes: sample mean=90
 sample variance=25

$$\begin{aligned} \square P(X|\text{Class}=\text{No}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$\begin{aligned} \square P(X|\text{Class}=\text{Yes}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

Since $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Therefore $P(\text{No}|X) > P(\text{Yes}|X)$
 $\Rightarrow \text{Class} = \text{No}$

Avoiding the 0-Probability Problem

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

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

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

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

c: number of classes

p: prior probability

m: parameter



Naïve Bayes (Summary)

- Advantage
 - Robust to isolated noise points
 - Handle missing values by ignoring the instance during probability estimate calculations
 - Robust to irrelevant attributes
- Disadvantage
 - Assumption: class conditional independence, which may cause loss of accuracy
 - Independence assumption may not hold for some attribute. Practically, dependencies exist among variables
 - Use other techniques such as Bayesian Belief Networks (BBN)



- Bayes' rule can be turned into a classifier
- Maximum A Posteriori (MAP) hypothesis estimation incorporates prior knowledge; Max Likelihood (ML) doesn't
- Naive Bayes Classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attributes are independent given the class.
- Bayesian classification is a generative approach to classification



Institut Informatika & Bisnis
DARMAJAYA
Yayasan Alfian Husin



**Kampus
Merdeka**
INDONESIA JAYA

**MERDEKA
BELAJAR**

THANK YOU!!

DATA SCIENCE DARMAJAYA "YOUR BEST FUTURE IN DATA"