MYONGJI UNIVERSITY

Artificial Intelligence
Lecture 7. Introduction to Classification

Spring 2022

Prof. Jonghoon Chun, Ph.D.

E-mail : jchun@mju.ac.kr
Lecture Note : http://lms.mju.ac.kr

Agenda

- Basic Concept
- Bayes Classification Methods

BASIC CONCEPT

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Classification vs. Numeric Prediction

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data

- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values

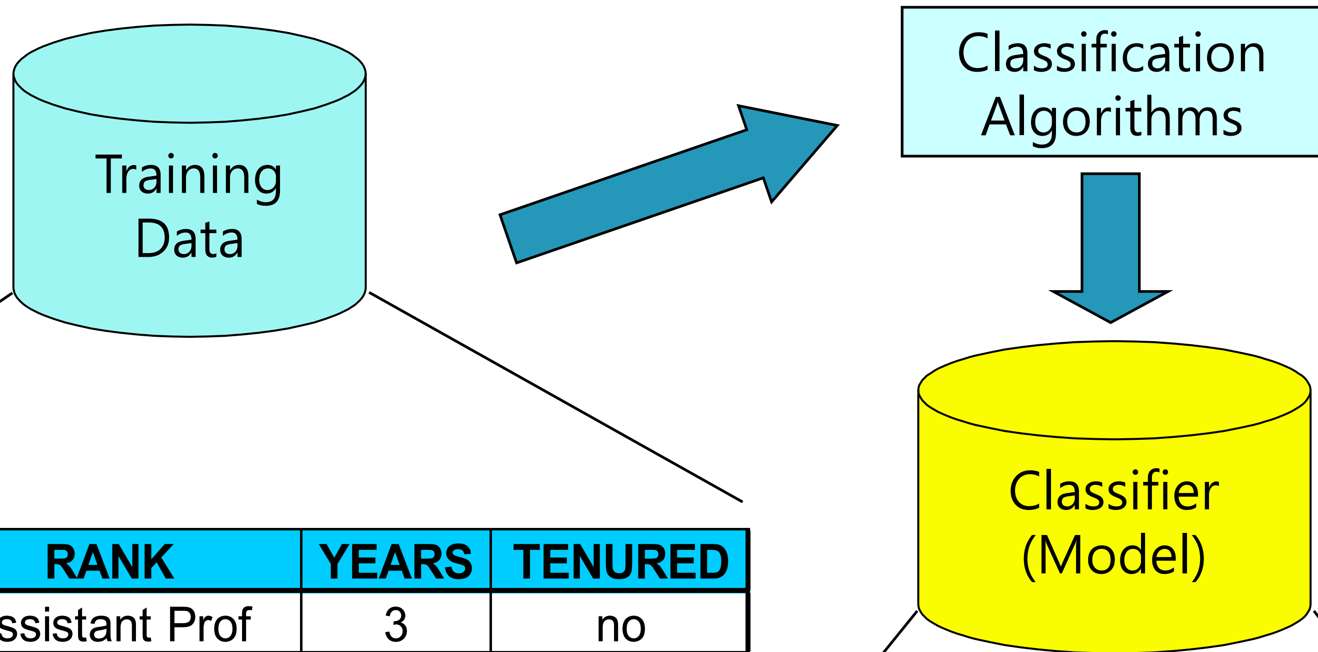
Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae

- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set (otherwise overfitting)
  - If the accuracy is acceptable, use the model to classify new data

- Note: If the test set is used to select models, it is called **validation (test) set!**

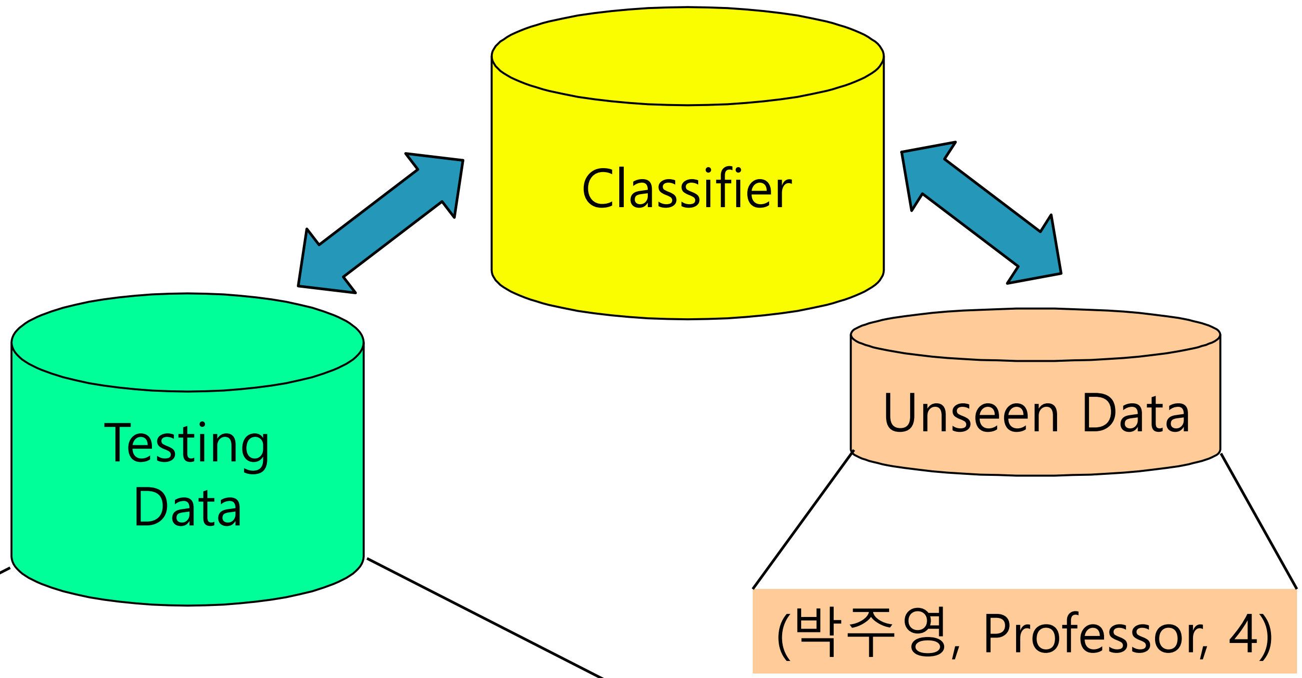
Process 1: Model Construction



NAME	RANK	YEARS	TENURED
김일주	Assistant Prof	3	no
이강선	Assistant Prof	7	yes
전종훈	Professor	2	yes
김상균	Associate Prof	7	yes
김대원	Assistant Prof	6	no
최성운	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Process 2: Use the Model in Prediction



NAME	RANK	YEARS	TENURED
신혜련	Assistant Prof	2	no
유형준	Associate Prof	7	no
임경훈	Professor	5	yes
윤석용	Assistant Prof	7	yes

Tenured?

YES!

BAYES CLASSIFICATION METHODS

Probabilistic Model

- Inference and conditional probabilities

	Preference: TV	Preference: Books	
female	1	2	1+2=3
male	4	3	4+3=7
	1+4=5	2+3=5	3+7=10 or 5+5=10

  - The probability a randomly sampled person in this group will be female is $P(female) = 3/10 = .3$
  - "Joint" probability
    - $P(female, books) = 2/10 = .2$
    - $P(x, y) = P(x \mid y) P(y)$
    - $P(female, books) = P(female \mid books) P(books) = 2/5 * 5/10 = .2$
    - $P(x \mid y) \geq P(x, y)$

But …

- $P(x, y) = P(x|y)\, P(y)$
- 즉 $P(x, y) \neq P(x)\, P(y)$
- BUT $P(x, y) = P(x)\, P(y)$ if x and y are statistically independent!

Baye's Theorem

- Baye's rule

$$P(x, y) = P(y, x)$$

$$P(x \mid y)\, P(y) = P(y \mid x)\, P(x)$$

$$P(x \mid y) = \frac{P(y \mid x)\, P(x)}{P(y)}$$

- compute conditional probabilities in terms of other probabilities
- Baye's rule may be thought of as describing evidence (e) and the relative degree of support it provides for a hypothesis (h)

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?

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

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables

- Given a record with attributes (X_1, X_2,…, X_d)
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes $P(Y \mid X_1, X_2,\ldots, X_d)$

- Can we estimate $P(Y \mid X_1, X_2,\ldots, X_d)$ directly from data?

Example Data

Given a Test Record:

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

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

- Can we estimate

  P("Evade = Yes" | X) and

  P("Evade = No" | X)?

  Replace

  "Evade = Yes" by Yes, and

  "Evade = No" by No

Using Bayes Theorem for Classification

- Approach:
  - Compute posterior probability $P(Y \mid X_1, X_2, \ldots, X_d)$ using the Bayes theorem

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

  - *Maximum a-posteriori* : Choose Y that maximizes
    $P(Y \mid X_1, X_2, \ldots, X_d)$
  - Equivalent to choosing value of Y that maximizes
    $P(X_1, X_2, \ldots, X_d \mid Y)\, P(Y)$

    since $P(X_1, X_2, \ldots, X_d)$ is constant for all classes.
- How to estimate $P(X_1, X_2, \ldots, X_d \mid Y)$?

Example Data

Given a Test Record:

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

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

Using Bayes Theorem:

- $P(\text{Yes} \mid X) = \dfrac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} \mid X) = \dfrac{P(X \mid \text{No})P(\text{No})}{P(X)}$
- How to estimate $P(X \mid \text{Yes})$ and $P(X \mid \text{No})$?

Naïve Bayes Classifier

- Assume independence among attributes X_i when class is given:
  - $P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$
  - Now we can estimate $P(X_i \mid Y_j)$ for all X_i and Y_j combinations from the training data
  - New point is classified to Y_j if $P(Y_j)\, \Pi\, P(X_i \mid Y_j)$ is maximal.

Naïve Bayes on Example Data

Given a Test Record:

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

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

- P(X | Yes) =
  P(Refund = No | Yes) x
  P(Divorced | Yes) x
  P(Income = 120K | Yes)

- P(X | No) =
  P(Refund = No | No) x
  P(Divorced | No) x
  P(Income = 120K | No)

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

- Class: $P(Y) = N_c/N$
  - e.g., $P(No) = 7/10$, $P(Yes) = 3/10$
- For categorical attributes:

$$P(X_i \mid Y_k) = |X_{ik}| / N_c$$

  - where $|X_{ik}|$ is number of instances having attribute value X_i and belonging to class Y_k
  - Examples:
    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes) = 0

Estimate Probabilities from Data – continuous value

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

- Gaussian distribution:

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (X_i, Y_j) pair

- 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:

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

Naïve 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) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
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 | No) = P(Refund=No | No)
  $\times$ P(Divorced | No)
  $\times$ P(Income=120K | No)
  = 4/7 $\times$ 1/7 $\times$ 0.0072 = 0.0006

- P(X | Yes) = P(Refund=No | Yes)
  $\times$ P(Divorced | Yes)
  $\times$ P(Income=120K | Yes)
  = 1 $\times$ 1/3 $\times$ 1.2 $\times$ 10^{-9} = 4 $\times$ 10^{-10}

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

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

=> Class = No!

Issues with Naïve Bayes Classifier

Naïve 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) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
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(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

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**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Naïve Bayes will not be able to classify X as Yes or No!

Issues with Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero
- Need to use other estimates of conditional probabilities than simple fractions
- Probability estimation:

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

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

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

c: number of classes

p: prior probability of the class

m: parameter

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

Example

- Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)

- Use **Laplacian correction** (or Laplacian estimator)
  - Adding 1 to each case

    P(income = low) = 1/1003

    P(income = medium) = 991/1003

    P(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts

Multinomial Naïve Bayes

- Naïve Bayes algorithm for multinomially distributed data
  - Multinomial data distribution models the probability of counts for rolling a *k*-sided die *n* times
  - a generalization of the binomial distribution (k = 2, n > 1)

- Frequently used in text classification
  - Document is represented as term vector counts (or tf-idf vectors)
  - The distribution is parametrized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the number of terms* (BOW size)
  - θ_{yi} is the probability $P(x_i|y)$ of term i appearing in a sample belonging to class y.

*term = feature = column = dimension = ...

Multinomial Naïve Bayes

- Probability Estimation for multinomial distribution

$$P(x_i \mid y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

  - Where N_{yi}: number of times term* i appears in a sample class y
  - N_y : total count of all terms for class y
  - n : number of terms*

- α : tuning parameter
  - α = 1: Laplace smoothing
  - α < 1: Lidstone smoothing

*term = feature = column = dimension = ...

Naïve Bayes Classifier Summary

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
    - E.g., hospitals: patients: Profile: age, family history, etc. Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier
- Need to use other techniques such as Bayesian Belief Networks (BBN)

END