CLASSIFICATION

Slides by:
Shree Jaswal

TOPICS TO BE COVERED

- Basic Concepts;
- **Classification methods**:

1. Decision Tree Induction: Attribute Selection Measures, Tree pruning.

2. Bayesian Classification: Naïve Bayes' classifier

- **Prediction**: Structure of regression models; Simple linear regression, Multiple linear regression.
- **Model Evaluation & Selection:** Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap; Comparing Classifier performance using ROC Curves.
- **Combining Classifiers**: Bagging, Boosting, Random Forests.

WHICH CHAPTER FROM WHICH TEXT BOOK ?

- Chapter 8: Classification: Basic Concepts from Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
- Chapter 4: Classification: Basic Concepts, decision trees and model evaluation from P. N. Tan, M. Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson Education

COURSE OUTCOME ADDRESSED

- **TEITC604.3**: Implement the appropriate data mining methods like classification, clustering or association mining on large data sets.
- **TEITC604.4**: Define and apply metrics to measure the performance of various data mining algorithms.
- **TEITC604.5**: Implement Prediction using Regression technique

CLASSIFICATION: DEFINITION

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

CLASSIFICATION

- Maps data into predefined groups or classes.
- Two step process
  - Training set
    - A model built describing a predetermined set of data classes
    - Supervised learning
  - Use model for classification
    - Accuracy of the model is first estimated.
    - Then classify/ predict the data.

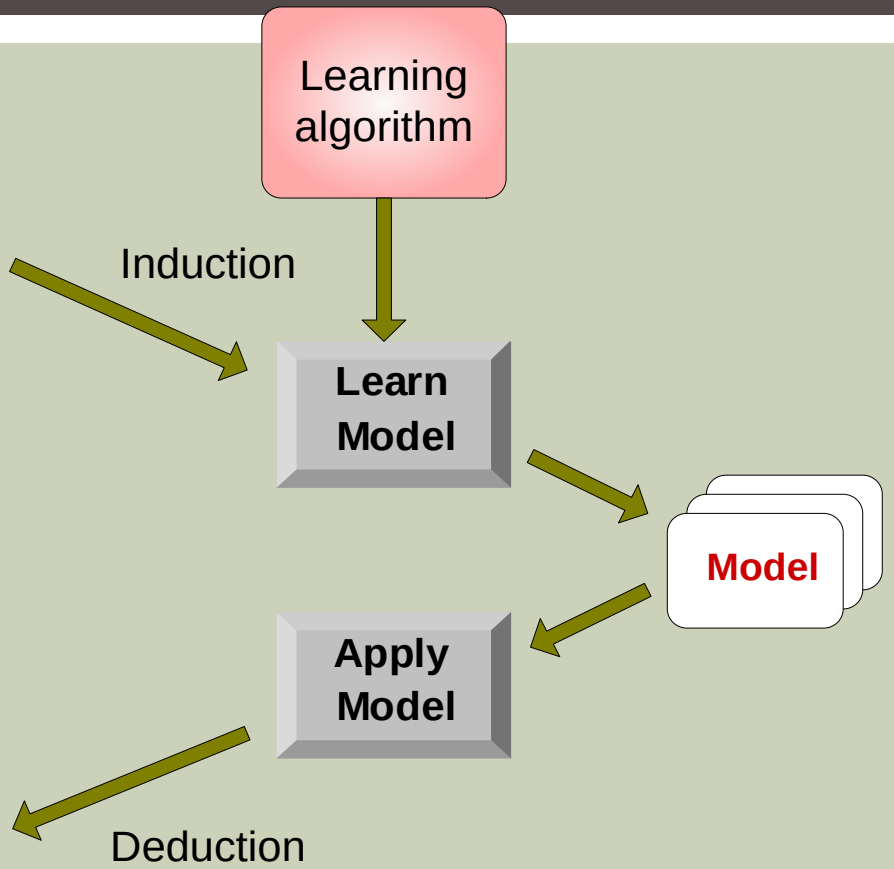
ILLUSTRATING CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

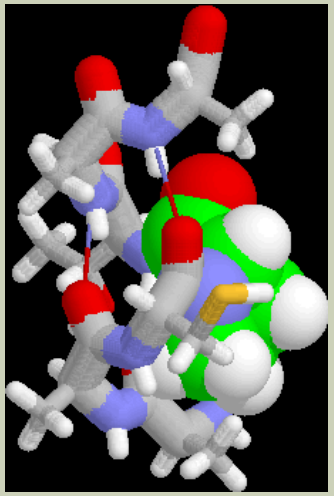
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



EXAMPLES OF CLASSIFICATION TASK

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc



CLASSIFICATION TECHNIQUES

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

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
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

BAYESIAN THEOREM: BASICS

- Let **X** be a data sample (“*evidence*”): class label is unknown
  - *E.g., X is a 35 year old customer with an income of ₹40,000*
- Let H be a *hypothesis* that X belongs to class C
  - *E.g., H is a hypothesis that our customer will buy a computer*
- P(H) (*prior probability*), the initial probability
  - *E.g., **X** will buy computer, regardless of age, income, …*
- Classification is to determine P(H|**X**), the probability that the hypothesis holds given the observed data sample **X**

BAYESIAN THEOREM: BASICS

- P(**X**|H) (*posteriori probability*), the probability of observing the sample **X**, given that the hypothesis holds
  - *E.g.,, the prob. that X is 35 years old and earns ₹40,000, given that we know **X** will buy computer*
- P(X): probability that sample data is observed
  - *E.g., Probability that a person from the dataset is 35 years old and earns ₹40,000*

BAYESIAN THEOREM

- Given training data **X**, *posteriori probability of a hypothesis* H, P(H|**X**), follows the Bayes theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

- Informally, this can be written as

  posteriori = likelihood x prior/evidence

- Predicts **X** belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|\mathbf{X})$ for all the *k* classes

- Practical difficulty: require initial knowledge of many probabilities, significant computational cost

TOWARDS NAÏVE BAYESIAN CLASSIFIER

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$
- Suppose there are *m* classes $C_1, C_2, \ldots, C_m$.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

NAÏVE BAYESIAN CLASSIFIER: TRAINING DATASET

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data sample
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

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

AVOIDING THE 0-PROBABILITY PROBLEM

- Naïve Bayesian prediction requires each conditional prob. be non-zero. Otherwise, the predicted prob. will be zero

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- Ex. Suppose a dataset with 1000 tuples,for class buys_computer=yes, income=low (0), income= medium (990), and income = high (10),
- Use Laplacian correction (or Laplacian estimator)
  - Adding 1 to each case

    Prob(income = low) = 1/1003

    Prob(income = medium) = 991/1003

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

NAÏVE BAYESIAN CLASSIFIER: COMMENTS

- 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 Bayesian Classifier
- How to deal with these dependencies?
  - Bayesian Belief Networks

EXAMPLE

Predict if Bob will default his loan

Bob

Home owner: *No*

Marital status: *Married*

Job experience: *3*

Home owner	Marital Status	Job experience (1-5)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes

Bob

Home owner: *No*

Marital status: *Married*

Job experience: *3*

Home owner	Marital Status	Job experience (1-5)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes

$$P(Y = No) = 7/10$$
$$P(Home\ owner = No | Y = No) = 4/7$$
$$P(Marital\ status = Married | Y = No) = 4/7$$
$$P(Job\ experience = 3 | Y = No) = 2/7$$

$$P(Bob\ will\ NOT\ default) = \frac{7}{10} \times \frac{4}{7} \times \frac{4}{7} \times \frac{2}{7} = \mathbf{0.065}$$

Bob

Home owner: *No*

Marital status: *Married*

Job experience: *3*

Home owner	Marital Status	Job experience (1-5)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes

$$P(Y = Yes) = 3/10$$
$$P(Home\ owner = No | Y = Yes) = 1/3$$
$$P(Marital\ status = Married | Y = Yes) = 1/3$$
$$P(Job\ experience = 3 | Y = Yes) = 1/3$$

$$P(Bob\ will\ default) = \frac{3}{10} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \mathbf{0.011}$$

Bob

Home owner: *No*

Marital status: *Married*

Job experience: *3*

$$P(Bob\ will\ NOT\ default) = \mathbf{0.065}$$
$$P(Bob\ will\ default) = \mathbf{0.011}$$

Predict: BOB WILL NOT DEFAULT

Home owner	Marital Status	Job experience (1-5)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes