OUTLINE

- Introduction to Bayesian Classification
  - Bayes Theorem
  - Naïve Bayes Classifier
  - Classification Example
- Text Classification – an Application
- Comparison with other classifiers
  - Advantages and disadvantages
  - Conclusions

CLASSIFICATION

- Classification:
  - predicts categorical class labels
  - 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

- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

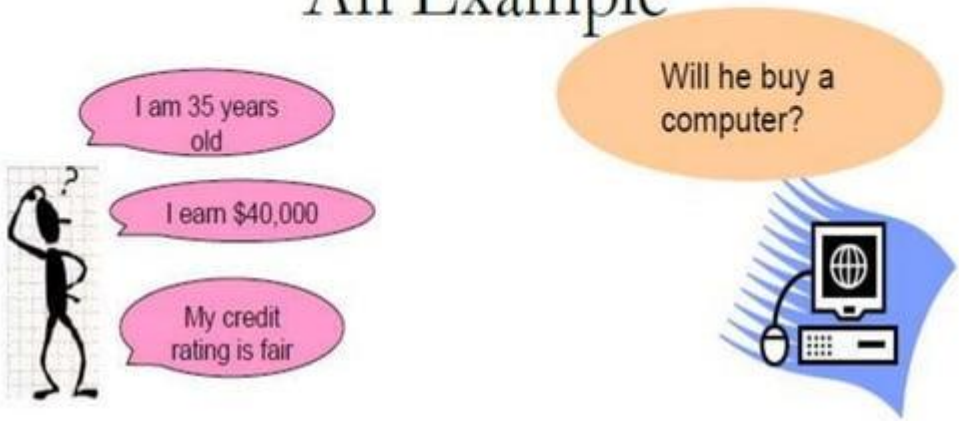
A TWO STEP PROCESS

- 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: 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 over-fitting will occur

Introduction to Bayesian Classification

- What is it ?
  - Statistical method for classification.
  - Supervised Learning Method.
  - Assumes an underlying probabilistic model, the Bayes theorem.
  - Can solve problems involving both categorical and continuous valued attributes.
  - Named after *Thomas Bayes, who proposed the Bayes Theorem.*

An Example



- X : 35 years old customer with an income of $40,000 and fair credit rating.
- H : Hypothesis that the customer will buy a computer.

The Bayes Theorem

- The Bayes Theorem:
  - $P(H|X) = P(X|H)\,P(H)/\,P(X)$

- $P(H|X)$: Probability that the customer will buy a computer given that we know his age, credit rating and income. (Posterior Probability of H)
- $P(H)$: Probability that the customer will buy a computer regardless of age, credit rating, income (Prior Probability of H)
- $P(X|H)$: Probability that the customer is 35 yrs old, have fair credit rating and earns $40,000, given that he has bought our computer (Posterior Probability of X)
- $P(X)$: Probability that a person from our set of customers is 35 yrs old, have fair credit rating and earns $40,000. (Prior Probability of X)

BAYESIAN CLASSIFIER

- D : Set of tuples
  - Each Tuple is an 'n' dimensional attribute vector
  - X : $(x_1, x_2, x_3, \ldots x_n)$
  - where x_i is the value of attribute A_i
- Let there are 'm' Classes : $C_1, C_2, C_3 \ldots C_m$
- Bayesian classifier predicts X belongs to Class C_i iff
  - $P(C_i|X) > P(C_j|X)$ for $1 <= j <= m$, $j \neq i$
- Maximum Posteriori Hypothesis
  - $P(C_i|X) = \dfrac{P(X|C_i)\,P(C_i)}{P(X)}$
  - Maximize $P(X|C_i)\,P(C_i)$ as $P(X)$ is constant

NAÏVE BAYESIAN CLASSIFIER...

- With many attributes, it is computationally expensive to evaluate $P(X|C_i)$
- Naïve Assumption of "class conditional independence"

$$P(X \mid C_i) = P(x_1, x_2, ..., x_n \mid C_i)$$

$$= P(x_1 \mid C_i) * P(x_2 \mid C_i) * ..... * P(x_n \mid C_i)$$

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

NAÏVE BAYESIAN CLASSIFIER...

To compute, $P(x_k|C_i)$

- A_k is categorical:

$$P(x_k|C_i) = \frac{\text{the number of tuples of class } C_i \text{ in } D \text{ having the value } x_k \text{ for } A_k}{\text{the number of tuples of class } C_i \text{ in } D.}$$

- A_k is continuous:

A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}).$$

"ZERO" PROBLEM

- What if there is a class, Ci, and X has an attribute value, xk, such that none of the samples in Ci has that attribute value?

- In that case P(xk|Ci) = 0, which results in P(X|Ci) = 0 even though P(xk|Ci) for all the other attributes in X may be large.

NUMERICAL UNDERFLOW

- When p(x|Y) is often a very small number: the probability of observing any particular high-dimensional vector is small.
- This can lead to numerical under flow.

$$P(X \mid C_i) = P(x_1, x_2, ..., x_n \mid C_i)$$

$$= P(x_1 \mid C_i) * P(x_2 \mid C_i) * ..... * P(x_n \mid C_i)$$

LOG SUM-EXP TRICK

$$c_{\text{map}} = \arg\max_{c \in \mathbf{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)].$$

BASIC ASSUMPTION

- The Naïve Bayes assumption is that all the features are conditionally independent given the class label.

$$P(X \mid C_i) = P(x_1, x_2, ..., x_n \mid C_i)$$

$$= P(x_1 \mid C_i) * P(x_2 \mid C_i) * ..... * P(x_n \mid C_i)$$

- Even though this is usually false (since features are usually dependent)

EXAMPLE: CLASS-LABELED TRAINING TUPLES FROM THE *CUSTOMER DATABASE*

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

EXAMPLE…

- X = (Age= '<=30', Income = medium, Student = yes, Credit_rating = fair)

- P(C1) = P(Buys_computer = yes) = 9/14 =0.643
- P(C2) = P(Buys_computer = no) = 5/14= 0.357

- $P(\text{Age}=\text{'<=30'} \mid \text{Buys\_computer} = \text{yes}) = \frac{\text{number of tuples with Buys\_computer= yes and Age <=30}}{\text{number of tuples with Buys\_computer = yes}}$

- P(Age='<=30' | Buys_computer = yes) = 2/9 =0.222

Similarly,
- P(Age='<=30' | Buys_computer = no) = 3/5 =0.600

- P(Income=medium | Buys_computer = yes) = 4/9 =0.444
- P(Income=medium | Buys_computer = no) = 2/5 =0.400

- P(Student=yes | Buys_computer = yes) = 6/9 =0.667
- P(Student=yes | Buys_computer = no) = 1/5 =0.200

- P(Credit_rating=fair | Buys_computer = yes) = 6/9 =0.667
- P(Credit_rating=fair | Buys_computer = no) = 2/5 =0.400

EXAMPLE…

- P(X | Buys a computer = yes)
  = P(Age='<=30' | buys_computer = yes) *
  P(Income=medium | buys_computer = yes) *
  P(Student=yes | buys_computer = yes) *
  P(Credit rating=fair | buys_computer = yes)
  = 0.222 * 0.444 * 0.667 * 0.667 = 0.044
- P(X | Buys a computer = No)
  = 0.600 * 0.400 * 0.200 * 0.400 = 0.019

- Find class C_i that Maximizes $P(X|C_i) * P(C_i)$
  →P(X | Buys a computer = yes) * P(Buys_computer = yes) = 0.028
  →P(X | Buys a computer = No) * P(Buys_computer = no) = 0.007

- Prediction : Buys a computer for Tuple X

USES OF NAÏVE BAYES CLASSIFICATION

- Text Classification
- Spam Filtering
- Hybrid Recommender System
  - Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource
- Online Application
  - Simple Emotion Modeling

Text Classification – An Application of Naive Bayes Classifier

WHY TEXT CLASSIFICATION?

- Learning which articles are of interest
- Classify web pages by topic
- Information extraction
- Internet filters

EXAMPLES OF TEXT CLASSIFICATION

- CLASSES=BINARY
  - "spam" / "not spam"
- CLASSES =TOPICS
  - "finance" / "sports" / "politics"
- CLASSES =OPINION
  - "like" / "hate" / "neutral"
- CLASSES =TOPICS
  - "AI" / "Theory" / "Graphics"
- CLASSES =AUTHOR
  - "Shakespeare" / "Marlowe" / "Ben Jonson"

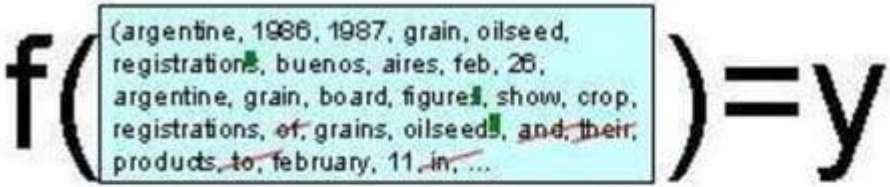
EXAMPLES OF TEXT CLASSIFICATION

- Classify news stories as world, business, SciTech, Sports ,Health etc
- Classify email as spam / not spam
- Classify business names by industry
- Classify email to tech stuff as Mac, windows etc
- Classify pdf files as research , other
- Classify movie reviews as favorable, unfavorable, neutral
- Classify documents
- Classify technical papers as Interesting, Uninteresting
- Classify Jokes as Funny, Not Funny
- Classify web sites of companies by Standard Industrial Classification (SIC)

Naïve Bayes Approach

- Build the Vocabulary as the list of all distinct words that appear in all the documents of the training set.
- Remove stop words and markings
- The words in the vocabulary become the attributes, assuming that classification is independent of the positions of the words
- Each document in the training set becomes a record with frequencies for each word in the Vocabulary.
- Train the classifier based on the training data set, by computing the prior probabilities for each class and attributes.
- Evaluate the results on Test data

REPRESENTING TEXT: A LIST OF WORDS



- Common Refinements: Remove Stop Words, Symbols

TEXT CLASSIFICATION ALGORITHM: NAÏVE BAYES

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

- Tct – Number of particular word in particular class
- Tct' – Number of total words in particular class
- B' – Number of distinct words in all class

EXAMPLE

► **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

EXAMPLE CONT…

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003.$$
$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001.$$

- Probability of Yes Class is more than that of No Class, Hence this document will go into Yes Class.

Advantages and Disadvantages of Naïve Bayes

Advantages :

- Easy to implement
- Requires a small amount of training data to estimate the parameters
- 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 modelled by Naïve Bayesian Classifier

An extension of Naive Bayes for delivering robust classifications

- Naive assumption (statistical independent. of the features given the class)
- NBC computes a single posterior distribution.
- However, the most probable class might depend on the chosen prior, especially on small data sets.
- Prior-dependent classifications might be weak.

Solution via set of probabilities:

- Robust Bayes Classifier (Ramoni and Sebastiani, 2001)
- Naive Credal Classifier (Zaffalon, 2001)

Relevant Issues

- Violation of Independence Assumption
- Zero conditional probability Problem

VIOLATION OF INDEPENDENCE ASSUMPTION

- Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naive."

IMPROVEMENT

- Bayesian belief network are graphical models, which unlike naïve Bayesian classifiers, allow the representation of dependencies among subsets of attributes.

- Bayesian belief networks can also be used for classification.

ZERO CONDITIONAL PROBABILITY PROBLEM

- If a given class and feature value never occur together in the training set then the frequency-based probability estimate will be zero.
- This is problematic since it will wipe out all information in the other probabilities when they are multiplied.
- It is therefore often desirable to incorporate a small-sample correction in all probability estimates such that no probability is ever set to be exactly zero.

CORRECTION

- To eliminate zeros, we use add-one or Laplace smoothing, which simply adds one to each count

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}},$$

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'},$$

EXAMPLE

- Suppose that for the class buys computer D (yes) in some training database, D, containing 1000 tuples.
  - we have 0 tuples with income D low,
  - 990 tuples with income D medium, and
  - 10 tuples with income D high.
- The probabilities of these events, without the Laplacian correction, are 0, 0.990 (from 990/1000), and 0.010 (from 10/1000), respectively.
- Using the Laplacian correction for the three quantities, we pretend that we have 1 more tuple for each income-value pair. In this way, we instead obtain the following probabilities :

$$\frac{1}{1003} = 0.001, \frac{991}{1003} = 0.988, \text{ and } \frac{11}{1003} = 0.011,$$

respectively. The "corrected" probability estimates are close to their "uncorrected" counterparts, yet the zero probability value is avoided.

Remarks on the Naive Bayesian Classifier

•Studies comparing classification algorithms have found that the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.

•Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Conclusions

✓The naive Bayes model is tremendously appealing because of its simplicity, elegance, and robustness.

✓ It is one of the oldest formal classification algorithms, and yet even in its simplest form it is often surprisingly effective.

✓It is widely used in areas such as text classification and spam filtering

✓A large number of modifications have been introduced, by the statistical, data mining, machine learning, and pattern recognition communities, in an attempt to make it more flexible.

✓ but some one has to recognize that such modifications are necessarily complications, which detract from its basic simplicity.

REFERENCES

- http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlbook/ch6.pdf
- Data Mining: Concepts and Techniques, 3rd Edition, Han & Kamber & Pei ISBN: 9780123 814791