ECLT5810/SEEM5750
Bayesian Classification

Prof. Wai Lam

Bayesian Classification (7.4)

- Bayesian classifiers are statistical classifiers.
- They predict class membership probabilities (probability that a given sample belongs to a particular class)
- *Probabilistic learning*:
  - Calculate explicit probabilities for hypothesis.
  - It is among the most practical approaches to certain types of learning problems
- Studies have shown that a simple Bayesian classifier known as the *naive Bayesian classifier* performs as well as decision tree and neural network classifiers.
- Bayesian classifiers are fast too.
- *Incremental*:
  - Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.

Bayes Theorem

- Bayesian classification is based on *Bayes Theorem*.
- Let X be a data sample whose class label is unknown. Let H be some hypothesis (X belongs to a class C).
- For classification problems, we want to determine
  - P(H|X) – *posterior probability*
- We can estimate P(H|X) from training data by Bayes theorem.

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

where P(H) is the *prior probability* of H.

P(X): probability that sample data is observed

Naïve (Simple) Bayes Classifier (I)

- Suppose that there are m classes, $C_1, \ldots, C_m$. Let X be a data sample to be classified.
- The simple Bayesian classifier assigns X to the class C_i if and only if

$$P(C_i \mid X) > P(C_j \mid X) \quad \text{for } 1 \leq j \leq m, j \neq i$$

The class C_i for which $P(C_i|X)$ is maximized is called the *maximum posterior hypothesis*.

Naïve (Simple) Bayes Classifier (II)

$$P(C_i \mid X) > P(C_j \mid X) \quad for\ 1 \le j \le m,\ j \ne i$$

Recall that

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need to be maximized. Therefore, X is assigned to the class C_i if and only if:

$$\mathrm{P}(\mathrm{X} \mid \mathrm{C_i})\mathrm{P}(\mathrm{C_i}) > \mathrm{P}(\mathrm{X} \mid \mathrm{C_j})\mathrm{P}(\mathrm{C_j}) \quad \mathrm{for}\ 1 \le \mathrm{j} \le \mathrm{m},\ \mathrm{j} \ne \mathrm{i}$$

Naïve (Simple) Bayes Classifier (III)

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \quad \text{for } 1 \le j \le m, j \ne i$$

$P(C_i)$ can be easily estimated.
For $P(X|C_i)$, we need to make the assumption of class conditional independence:

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

The probabilities $P(X_k|C_i)$ can be easily estimated from training samples

“Buys Computer” Dataset

Class:
C1:buys_computer=‘yes’
C2:buys_computer=‘no’

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…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

Prediction: X =(age<=30, income=medium, student=yes, credit_rating=fair)

Naïve Bayesian Classifier: Example

Training Stage (Compute P(X/Ci) 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

Prediction/Testing Stage:

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

P(X|Ci) : P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.0.667 =0.044
P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019
P(X|Ci)*P(Ci) : P(X|buys_computer="yes") * P(buys_computer="yes")=0.028
P(X|buys_computer="no") * P(buys_computer="no")=0.007

Then, X is predicted to belong to the class "buys_computer=yes"

The "zero-frequency problem"

- What if an attribute value doesn't occur with every class value?
  - Probability will be zero!

- A posteriori probability will also be zero!
(No matter how likely the other values are!)

  $\Pr[class|X] = 0$

- Remedy: add 1 to the count for every attribute value-class combination (Laplace estimator)
- Result: probabilities will never be zero!
  - also: stabilizes probability estimates

With add-1 Adjustment

Compute P(X/Ci) for each class

P(age="<=30" | buys_computer="yes")

= P(age="<=30", buys_computer="yes") / P(buys_computer="yes")

= P(age="<=30", buys_computer="yes") /

[P(age="<=30", buys_computer="yes") +

P(age="31..40", buys_computer="yes") +

P(age=">40", buys_computer="yes")]

= (2+1)/(9+3)

P(age="<=30" | buys_computer="no") = (3+1)/(5+3)

P(income="medium" | buys_computer="yes")= (4+1)/(9+3)

P(income="medium" | buys_computer="no") = (2+1)/(5+3)

P(student="yes" | buys_computer="yes)= (6+1)/(9+2)

P(student="yes" | buys_computer="no")= (1+1)/(5+2)

P(credit_rating="fair" | buys_computer="yes")=(6+1)/(9+2)

P(credit_rating="fair" | buys_computer="no")=(2+1)/(5+2)

Naïve Bayesian Classifier: Comments

- Advantages :
  - Easy to implement
  - Good results obtained in most of the cases
  - Nevertheless, the performance of Bayesian classifiers is comparable to decision tree and neural network in some domains.
- 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* (out of our scope)