Knowledge Discovery & Data Mining
— Classification: Bayesian Classification —

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Outline

- Bayesian Classification
  - Bayes’ Theorem, posterior, likelihood, prior, and marginal probability
  - Prediction Based on Bayes’ Theorem
  - Naïve Bayes Classifier

Bayesian Classification: Why?

- **A statistical classifier**: performs probabilistic prediction, i.e., predicts class membership probabilities
- **Theoretical 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 sample can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

Bayes' Theorem: Basics

Named after: Thomas Bayes, an 18th-century English clergyman, who did early work in probability and decision theory.

Consider X as a data tuple. Within Bayesian context, X is viewed as "evidence." Typically, this evidence is characterized by measurements across a set of n attributes. Let's define H as a hypothesis suggesting that this data tuple, X, belongs to a specific class C. For classification tasks, our aim is to determine $P(H|X)$, which represents the probability of hypothesis H being true based on the observed evidence X. Essentially, we're trying to assess the likelihood of X being in class C, given its attribute composition.

Bayes' Theorem: Basics

- **P(H|X)**: Posterior probability (probability tuple X belongs to class given its attributes).
  - the probability that customer X will buy a computer given that we know the customer's age and income.
- **P(H)**: Prior probability (probability of a hypothesis without evidence).
  - the probability that any given customer will buy a computer, regardless of age, income, or any other information
- **P(X|H)**: Likelihood (probability of evidence given a hypothesis).
  - if we know a customer will buy a computer, what is the probability that this customer X is 35 years old and earns $40,000?
- **P(X)**: Marginal probability (probability of X).
  - the probability that a person from our set of customers is 35 years old and earns $40,000.

Bayes' theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X|H), and P(X). Bayes' theorem is

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

Prediction Based on Bayes' Theorem

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

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

- Informally, this can be viewed 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 $P(C_k|X)$ for all the *k* classes
- Practical difficulty: It requires initial knowledge of many probabilities, involving significant computational cost.

Classification Is to Derive the Maximum Posteriori

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **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

Challenge: *Estimating P(X|Ci) is challenging due to the exponential attribute value space.*

Naïve Bayes Classifier

The **Naïve Bayesian** classifier, or simple Bayesian classifier, follows the same procedure as Bayes classifier, except the way it estimates the conditional probabilities. In detail, it works as follows:

1. Training Data Representation:

- Let **D** be the training set containing tuples and their corresponding class labels.
- Every tuple is depicted by an n-dimensional attribute vector: **X = (x_1, x_2, ... , x_n)**
- Here, **X** describes n measurements from attributes **A1, A2, ... , An** respectively.

2. Class Prediction:

- If we have m classes, represented as $C_1, C_2, \ldots, C_m$, the classifier predicts the class of tuple X based on the highest posterior probability.
- The formula is represented by Bayes' theorem:

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

- The class C_i for which $P(C_i|X)$ is maximized is termed the **maximum posteriori hypothesis**.

Naïve Bayes Classifier

3. Computing Posterior Probability:

- Given that $P(X)$ is consistent across all classes, the main goal is to identify the class that maximizes $P(X|C_i)P(C_i)$.
- If class prior probabilities are unknown, classes are typically assumed to be equally likely, which means the focus is on maximizing $P(X|C_i)$.
- Alternatively, you can estimate class prior probabilities as:

$$P(C_i) = \frac{|C_i, D|}{|D|}$$

where $|C_i, D|$ is the number of training tuples of class C_i in D.

4. The Naïve Assumption:

- Computing $P(X|C_i)$ with multiple attributes can be computationally intense.
- The naïve assumption of class-conditional independence is made to simplify computation, meaning attributes are considered independent given a class label.

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \end{aligned}$$

where x_k represents the value of attribute A_k for tuple X.

Naïve Bayes Classifier

5. Categorical vs. Continuous Attributes:

- For categorical attribute A_k, $P(x_k|C_i)$ is determined by:

$$P(x_k|C_i) = \frac{\text{Number of tuples of class } C_i \text{ with value } x_k \text{ for } A_k}{|C_i, D|}$$

- For continuous attributes, a Gaussian distribution is often assumed with mean (μ) and standard deviation (σ):

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

where

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

For example, considering attributes age and income, if customers who buy a computer have an average age of 38 with a standard deviation of 12, we can use the above formula to estimate the probability for a given age.

6. Prediction:

- For class label prediction of X, $P(X|C_i)P(C_i)$ is evaluated for each class.
- The predicted class label for X is the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

Naïve Bayes Classifier

Example. Naïve Bayesian Classification for Predicting a Class Label. Given the following training set, D. and a new tuple. X = (age = youth, income = medium, student = yes, credit-rating = fair), our goal is to predict its class label using the naïve Bayesian classification method.

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

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

$$P(C_i) = \frac{|C_i, D|}{|D|}$$

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

$$= P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

$$P(x_k|C_i) = \frac{\text{Number of tuples of class } C_i \text{ with value } x_k \text{ for } A_k}{|C_i, D|}$$

Naïve Bayes Classifier

Example. Naïve Bayesian Classification for Predicting a Class Label. Given the following training set, D. and a new tuple. X = (age = youth, income = medium, student = yes, credit-rating = fair), our goal is to predict its class label using the naïve Bayesian classification method.

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

Prior Probabilities:

1. $P(buys\_computer = yes) = \frac{9}{14} = 0.643$
2. $P(buys\_computer = no) = \frac{5}{14} = 0.357$

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

$$P(C_i) = \frac{|C_i, D|}{|D|}$$

Naïve Bayes Classifier

Computing Probabilities for Given Tuple:

1. $P(X|buys\_computer = yes) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
2. $P(X|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$

RID	age	income	student	credit_rating	Class:
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

Conditional Probabilities:

1. $P(age = youth|buys\_computer = yes) = \frac{2}{9} = 0.222$
2. $P(age = youth|buys\_computer = no) = \frac{3}{5} = 0.600$
3. $P(income = medium|buys\_computer = yes) = \frac{4}{9} = 0.444$
4. $P(income = medium|buys\_computer = no) = \frac{2}{5} = 0.400$
5. $P(student = yes|buys\_computer = yes) = \frac{6}{9} = 0.667$
6. $P(student = yes|buys\_computer = no) = \frac{1}{5} = 0.200$
7. $P(credit\_rating = fair|buys\_computer = yes) = \frac{6}{9} = 0.667$
8. $P(credit\_rating = fair|buys\_computer = no) = \frac{2}{5} = 0.400$

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

$$P(C_i) = \frac{|C_i, D|}{|D|}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

$$P(x_k|C_i) = \frac{\text{Number of tuples of class } C_i \text{ with value } x_k \text{ for } A_k}{|C_i, D|}$$

Naïve Bayes Classifier

Example. Naïve Baye … Giv … h,
income = medium, s … edi … .

Prior Probabilities:

1. $P(buys\_computer = yes) = \frac{9}{14} = 0.643$
2. $P(buys\_computer = no) = \frac{5}{14} = 0.357$

RID	age	income	student	credit_rating	Class: buys
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

Computing Probabilities for Given Tuple:

1. $P(X|buys\_computer = yes) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
2. $P(X|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$

Class Maximization:

1. $P(X|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$
2. $P(X|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$

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

$$P(C_i) = \frac{|C_i, D|}{|D|}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

$$P(x_k|C_i) = \frac{\text{Number of tuples of class } C_i \text{ with value } x_k \text{ for } A_k}{|C_i, D|}$$

Avoiding the Zero-Probability Problem

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

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

- Ex. 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*
  - 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 Bayes Classifier: Advantages vs. Disadvantages

- Advantages
  - Simple and easy to implement.
  - Provides good results in many scenarios, especially with large datasets.
- Disadvantages
  - Naïve Bayes assumes that features are conditionally independent given the class label, which can lead to a loss in accuracy when dependencies exist.
  - In practical applications, dependencies often exist between features that Naïve Bayes cannot capture. For instance, In a healthcare setting, features might include:
    - Patient Profile: age, family history, etc.   Symptoms: fever, cough, etc. Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier
- How to deal with these dependencies? Bayesian Belief Networks

Summary

- Bayesian Classification
  - Bayes' Theorem, posterior, likelihood, prior, and marginal probability
  - Prediction Based on Bayes' Theorem
  - Naïve Bayes Classifier