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 predictions,

i.e., predicts class membership probabilities based on observed data.

Why Bayesian Classification?

- Based on Bayes' Theorem. Useful in contexts with known prior probabilities and updating beliefs with new data.
- **Performance:**
  - Naïve Bayes (simplified Bayesian classifier) often rivals complex models (like decision trees and neural networks) despite its assumptions.
  - Fast, interpretable, and effective on a wide range of tasks.
- **Incremental**:
  - Efficiently updates with each new instance; allows for continuous learning without retraining.
  - Integrates prior knowledge, improving predictions with added samples.

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

- Objective: Use Bayes' theorem to classify a data point by determining the most probable class.
- **Problem Setup**
  - **Dataset** D: Consists of tuples (data points) with associated class labels.
  - **Attribute Vector** $X = (x_1, x_2, \ldots, x_n)$: Represents a data point with n features.
  - **Classes** $C_1, C_2, \ldots, C_m$: Define the possible categories for classification.
- **Goal**: Find the class C_i that maximizes the posterior probability $P(C_i|X)$, known as Maximum A Posteriori (MAP) estimation.
- By **Bayes' Theorem**:

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

- Since $P(X)$ is constant across classes, it can be ignored in maximization, reducing the goal to:

$$\text{Maximize } P(C_i|X) = P(X|C_i) \cdot P(C_i)$$

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, a probabilistic classifier based on Bayes' theorem with a strong independence assumption between features.

Applications: Widely used in spam detection, document classification, and medical diagnosis.

Advantages

- Assumes features contribute independently to the classification, which simplifies calculations.
- Fast and efficient on large datasets.
- Handles both categorical and continuous data well with different approaches.

Naïve Bayes Classifier

Goal: Classify a new data point X by **maximizing the posterior probability** $P(C_i|X)$.

By **Bayes' Theorem**:

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

Since $P(X)$ is constant across classes, it can be ignored in maximization, reducing the goal to:

$$\text{Maximize } P(C_i|X) = \underline{P(X|C_i)} \cdot P(C_i)$$

Challenging !

The Naïve Assumption

Assumes **independence** between features, simplifying to:

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

Naïve Bayes Classifier

$$\text{Maximize } P(C_i|X) = \underline{P(X|C_i)} \cdot P(C_i)$$

The Naïve Assumption

Assumes **independence** between features, simplifying to:

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

Categorical attributes:

$$P(x_k|C_i) = \frac{\text{Count of } x_k \text{ in class } C_i}{|C_{i,D}|}$$

Continuous attributes:

Assumes a Gaussian (Normal) distribution:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi\sigma_{C_i}^2}} e^{-\frac{(x_k - \mu_{C_i})^2}{2\sigma_{C_i}^2}}$$

Naïve Bayes Classifier

$$\text{Maximize } P(C_i|X) = P(X|C_i) \cdot P(C_i)$$

Class Prior Probability $P(C_i)$:

- Estimated as:

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

where $|C_i, D|$ is the count of instances in class C_i and $|D|$ is the total number of instances.

Prediction with Naïve Bayes

$$\text{Maximize } P(C_i|X) = P(X|C_i) \cdot P(C_i)$$

For a New Instance X:

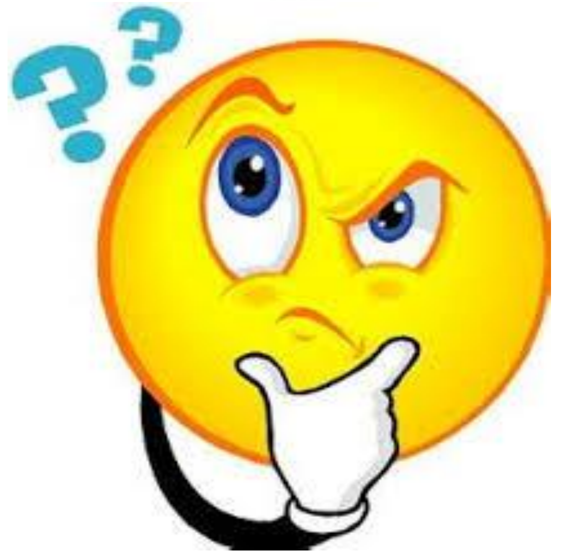
1. Calculate $P(X|C_i) \cdot P(C_i)$ for each class C_i.
2. **Prediction**: Assign X the class label C_i with the **highest posterior probability** $P(X|C_i) \cdot P(C_i)$.

Formula Recap:

$$\text{Predicted class for } X = \arg\max_{C_i} \; P(X|C_i) \cdot P(C_i)$$

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