Naïve Bayes

Lương Thái Lê

Course Outline

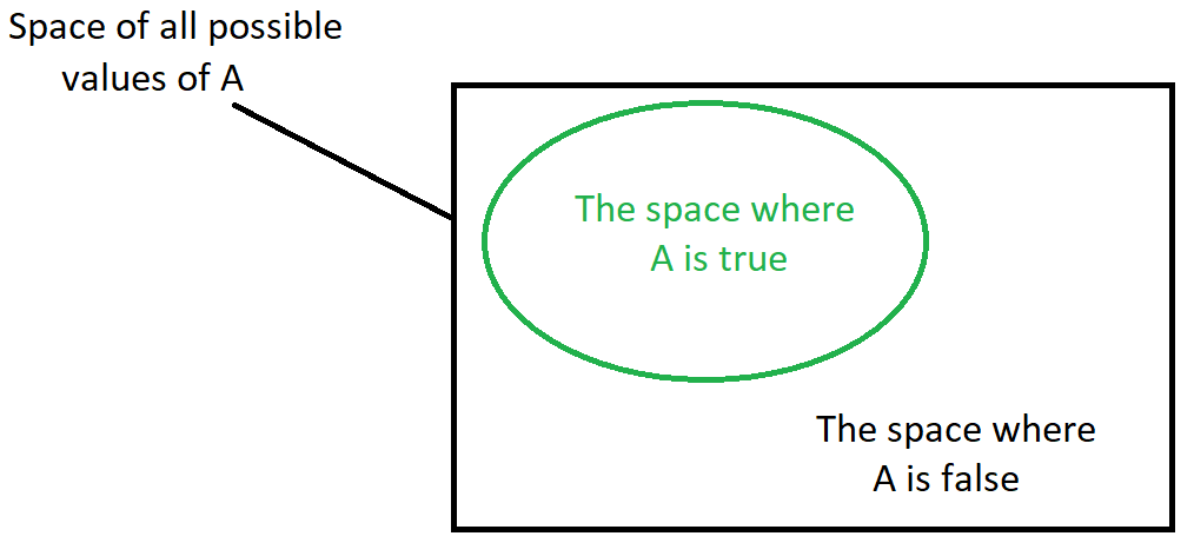
1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Basic concepts of probability

- Suppose we have an experiment (e.g. rolling a dice) whose outcome is random (depends on the probability)
- ***Space of possibilities S***: set of all possible outcomes
  - Eg: S = {1,2,3,4,5,6} for the dice roll experiment
- ***An event E***: a subset of S
  - Eg: E = {1,3,5}
- ***Event space W***: P(S)
- ***Random variable A***: a function represents an event, and there is a degree of probability that this event will occur.
  - Eg: A= “The number of dots is odd when the dice are rolled”

Probability representation

- P(A) = The part of the space of events (W) where A is true



Some Properties

- $0 \leq P(A) \leq 1$
- P(not A) = 1 − P(A)
- P(A V B)= P(A)+ P(B)−P(A ∧B)
- Suppose random variables A and B can take one of k (>2) values $\{v_1, v_2, \ldots, v_k\}$, then:
  - $P(A = v_i \wedge A = v_j) = 0$ if $i \neq j$
  - $P(A=v_1 \text{ V } A=v_2 \text{ V } \ldots \text{ V } A=v_k) = \sum_{i=1}^{k} P(A = v_i) = 1$
  - $P(B \wedge [A=v_1 \text{ V } A=v_2 \text{ V } \ldots \text{ V } A=v_m] = \sum_{i=1}^{m} (P(B \wedge A = v_i)$

Conditional Probability

- P(A|B) is the part of space W in which A is true, provided that B is true
  - Eg:

    A: “I will play football tomorow”

    B: “It won’t rain tomorow”

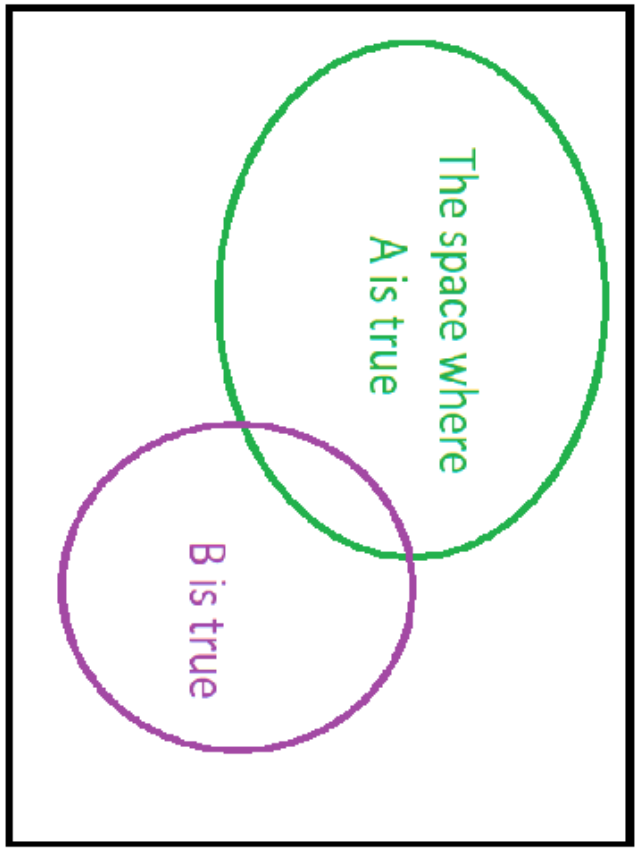
  => P(A|B) is the probability that I will play football tomorrow if it won’t rain

- Let P(A ∧ B) = P(A,B), then

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

P(A|B) + P(~A|B) = 1

$$\sum_{i=1}^{k} P(A = v_i \mid B) = 1$$



Probability independent variables (1)

- Two events (random variable) A and B are said to be probabilistically independent if the probability of event A is the same for all cases:
  - B happens
  - B does not happen
- Eg:
  - A: I will go swimming tomorrow
  - B: Long will go swimming tomorrow

$$P(A|B) = P(A)$$

Probability independent variables (2)

- $P(\sim A \mid B) = P(\sim A)$
- $P(B \mid A) = P(B)$
- $P(A, B) = P(A) . P(B)$
- $P(\sim A, B) = P(\sim A) . P(B)$
- $P(A, \sim B) = P(A) . P(\sim B)$
- $P(\sim A, \sim B) = P(\sim A) . P(\sim B)$

Probability independent variables with >2 variables

- P(A|B,C) is probability of A when B,C (is known)
- Two variables A and C are said to be conditionally independent of variable B, if the probability of A with respect to B is equal to the probability of A with respect to B and C.

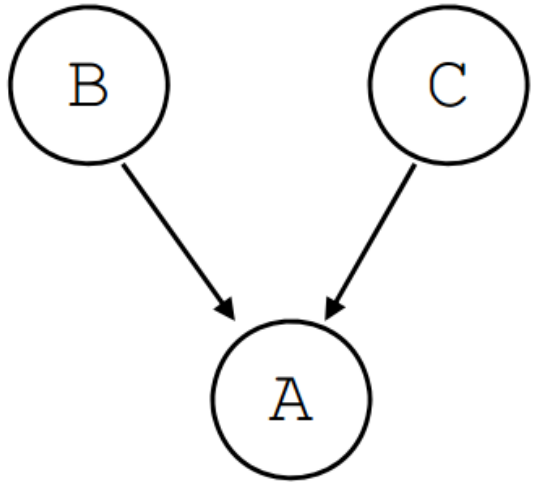
$$P(A|B,C) = P(A|B)$$

- Eg:

  A: I will play football tomorrow

  B: The football match will take place indoors

  C: It will rain tomorow



$P(A|B,C)$

Important Rules of Probability

- Chain rules:
  - $P(A,B) = P(A|B)\ P(B) = P(B|A)\ P(A)$
  - $P(A|B) = P(A,B)/P(B) = P(B|A).P(A)/P(B)$
  - $P(A,B|C) = P(A,B,C)/P(C) = P(A|B,C).P(B,C)/P(C) = P(A|B,C).P(B|C)$
- Probability independence and conditional independence
  - $P(A|B) = P(A)$; if A and B are probability independence
  - $P(A,B|C) = P(A|C).P(B|C)$; if A and B are conditional independence with C
  - $P(A_1,\ldots,A_n|C) = P(A_1|C)\ldots P(A_n|C)$; if A_i are conditional independence with C

Course Outline

1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Bayes Theorem

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

- `P(h)` : the prior probability of the (categorical) hypothesis h
- `P(D)` : the prior probability of the observing of data
- `P(D|h)` : the conditional probability of observing of data D, if hypothesis (category) h is known to be true
- **P(h|D): the conditional probability of the (categorical) hypothesis h being true, if the data D is observed**

=>**Probabilistic classification methods will use this conditional probability (called posterior probability)**

Bayes Theorem – Example (1)

- Suppose we have this data set

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes

Bayes Theorem – Example (2)

- **Data D**: *Outlook is Sunny and Wind is Strong*
- **Categorical hypothesis h**: He plays tennis
- **The prior probability P(h)**: Probability that he plays tennis (no matter how outdoors and windy)
- **The prior probability P(D)**: Probability that Outdoors is sunny and Wind is strong
- **P(D|h)**: Probability that Outdoors is sunny and Wind is strong, if he plays tennis
- **P(h|D)**: Probability that he plays tennis if Outdoors is sunny and Wind is strong

Maximum a Posteriori – MAP

- Given a set of possible hypotheses (target classes) H, the learning system will find the most probable hypothesis h(∈H) for the observed data D
- **This hypothesis h is called the maximum posterior**

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$h_{MAP} = \arg\max_{h \in H} \frac{P(D \mid h).P(h)}{P(D)}$$ (Bayes theorem)

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h).P(h)$$ (P(D) is the same with all h)

MAP - Example

- The set H includes 2 hypotheses:
  - h_1 : He plays tennis
  - h_2 : He dose not play tennis
- Calculate the 2 conditional probabilities: $P(h_1|D)$, $P(h_2|D)$
- if $P(h_1|D) > P(h_2|D)$ $h_{MAP} = h_1$
  else $h_{MAP} = h_1$
- Because $P(D) = P(D, h_1) + P(D, h_2)$ is the same for both hypotheses h_1 and h_2, so P(D) can be ignored.
- So if $P(D|h_1) . P(h_1) > P(D|h2) . P(h_2)$ then he plays tennis else he dosen't

Maximum Likelihood Estimation (MLE)

- Suppose that all assumptions have the same prior probability value: $P(h_i)=P(h_j)$, $\forall h_i,h_j \in H$
- The MLE method finds the hypothesis that maximizes the value $P(D|h)$; where $P(D|h)$ is called the likelihood of data D for h
- The maximum likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

MLE - Example

- The set H includes 2 hypothesis:
  - h_1 : He plays tennis
  - h_2 : He dose not play tennis

  D: Data set (dates) in which the Outlook attribute is Sunny and the Wind is Strong
- Find $P(D|h_1)$ and $P(D|h2)$:
  - $P(Outlook=Sunny,Wind=Weak|h_1)=1/8$
  - $P(Outlook=Sunny,Wind=Weak|h_2)=1/4$

$\Rightarrow h_{MLE} = h2 \Rightarrow$ He dose not play tennis

Course Outline

1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Naïve Bayes Classification – Idea (1)

- Classification problem:
  - Input:
    - A training data set D ={($x^{(i)}$,$c^{(j)}$)}; i= 1,…,m; j= 1,..,k where :
      - x: is an training example, is represented by n dimension vector (x_1, x_2, …, x_n)
      - c: is a label class in the target class set C = {c_1, c_2, …, c_k}
    - A test example z not belong to D
  - Output:
    - A classification model F
    - The class that z is determined belong to by F
- Motivation:
  - find c_{MAP} is the most suitable class for z

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z)$$

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z_1, z_2, \ldots, z_n)$$

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(z_1, z_2, \ldots, z_n \mid c_i).P(c_i)}{P(z_1, z_2, \ldots, z_n)}$$

Naïve Bayes

Naïve Bayes Classification – Idea (2)

- Because $P(z_1, z_2, \ldots, z_n)$ is the same for all class c_i , so we find:

$$c_{MAP} = argmax_{c_i \in C} P(z_1, z_2, \ldots, z_n | c_i) P(c_i)$$

- Assumption in the Naïve Bayes classifier, attributes are conditionally independent of classes:

$$P(z_1, z_2, \ldots, z_n | c_i) = \prod_{j=1}^{n} P(z_j | c_i)$$

- Naïve Bayes find the most likelihood class for z:

$$c_{NB} = argmax_{c_i \in C} P(c_i) \prod_{j=1}^{n} P(z_j | c_i)$$

Naïve Bayes Classification – Alg

- Training phase: with a training example x = (x_1, x_2, …, x_n)
  - Calculate P(c_i) for each class $c_i \in C$
  - Calculate P($x_j | c_i$) for each attribute $x_j \in$ vector x
- Classification phase: for a new example z = (z_1, z_2, …, z_n)
  - Calculate $P(c_i) \prod_{j=1}^{n} P(z_j|c_i)$
  - Find the most suitable class c* for z:

$$c^* = argmax_{c_i \in C} P(c_i) \prod_{j=1}^{n} P(z_j|c_i)$$

Course Outline

1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Naïve Bayes Classification – Example (1)

A young student with an average income and a normal credit rating would buy a calculator?

Rec. ID	Age	Income	Student	Credit_Rating	Buy_Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

Naïve Bayes Classification – Example (2)

- Problem modeling:
  - z = (Age= `Young`, Income= `Medium`, Student= `Yes`, Credit_Rating= `Fair`)
  - There are 2 class: c_1 (Bye computer); c_2 (Not bye computer)
- Calculate priorities
  - $P(c_1) = 9/14$
  - $P(c_2) = 5/14$
- Calculate the probability value of each attribute value for each subclass

$P(\text{Age}=\texttt{Young}|c_1) = 2/9$;

$P(\text{Income}=\texttt{Medium}|c_1) = 4/9$;

$P(\text{Student}=\texttt{Yes}|c_1) = 6/9$;

$P(\text{Credit\_Rating}=\texttt{Fair}|c_1) = 6/9$;

$P(\text{Age}=\texttt{Young}|c_2) = 3/5$

$P(\text{Income}=\texttt{Medium}|c_2) = 2/5$

$P(\text{Student}=\texttt{Yes}|c_2) = 1/5$

$P(\text{Credit\_Rating}=\texttt{Fair}|c_2) = 2/5$

Naïve Bayes Classification – Example (3)

- Calculate the probability (likelihood) of the example z for each class c_i

$$P(z|c_1) = P(\text{Age=Young}|c_1).P(\text{Income=Medium}|c_1).P(\text{Student=Yes}|c_1).$$
$$P(\text{Credit\_Rating=Fair}|c_1) = (2/9).(4/9).(6/9).(6/9) = 0.044$$

$$P(z|c_2) = P(\text{Age=Young}|c_2).P(\text{Income=Medium}|c_2).P(\text{Student=Yes}|c_2).$$
$$P(\text{Credit\_Rating=Fair}|c_2) = (3/5).(2/5).(1/5).(2/5) = 0.019$$

- Determine the most possible class
  - $P(c_1)P(z|c_1) = (9/14).0{,}044 = 0{,}028$
  - $P(c_2)P(z|c_2) = (5/14).\ 0{,}019 = 0{,}007$

=> Conclusion: He (z) will buy a computer

Course Outline

1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Naïve Bayes Classification – Problems (1)

- If there are no examples associated with class c_i having attribute value z_j :

$$P(z_j|c_i) = 0 \quad \Rightarrow \quad P(c_i)\prod_{j=1}^{n} P(z_j|c_i) = 0$$

- Solution: Using Bayes to estimate $P(z_j|c_i)$:

$$P(z_j|c_i) = \frac{n(c_i, z_j) + mp}{n(c_i) + m}$$

  - $n(c_i)$= number of training examples associated with c_i
  - $n(c_i, z_j)$ = number of training examples associated with c_i having attribute value z_j
  - p: estimate for the probability value $P(z_j|c_i)$

    => p = 1/k for feature f_j has k possible values
  - m: a weight is chosen

Naïve Bayes Classification – Problems (2)

- Limits on accuracy in computer calculations
  - $P(z_j|c_i)$<1, so if the number of attribute values is big then:

$$\lim(\prod_{j=1}^{n} P(z_j|c_i)) = 0$$

- Solution: Using the logarithmic function for probability values

$$c_{NB} = \arg\max_{c_i \in C} \left(\log \left[P(c_i) . \prod_{j=1}^{n} P(x_j \mid c_i) \right] \right)$$

$$c_{NB} = \arg\max_{c_i \in C} \left(\log P(c_i) + \sum_{j=1}^{n} \log P(x_j \mid c_i) \right)$$

Course Outline

1. Introduction to probability
2. Naïve Bayes Theorem
3. Naïve Bayes Algorithm
4. Naïve Bayes Examples
5. Naïve Bayes Problems
6. Naïve Bayes Conclusion

Naïve Bayes Classification – Conclusion

- One of the most commonly used machine learning methods in reality
- Despite assuming the conditional independence of the attributes for the classifiers, the Naïve Bayes classifier still obtains good classification results in many fields of practical applications.
- When to use?
  - Traing data set has large or medium size
  - Examples are represented by a large number of attributes
  - Attributes are conditionally independent for target classes