Naïve Bayes

Introduction

- Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem.
- Bayes theorem is the one that has assumptions on conditional independence.
- Used often in text mining – spam detection, document classification, mail sorting and so on.
- Computationally intensive.
- Needs limited training data, thus saves the training time.

Let us start with what Bayes theorem is -

Whos who is bayes??

- Bayes rule:

$$p(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

In Bayesian,
d is considered the evidence/ data.
h is the hypothesis.
In classification we want to determine

P(h|d), the probability that the hypothesis holds given d(observed data or some evidence),

Other way round, we are looking for probability that d belongs to C, given description of d.

Whos who is bayes??

- The central idea behind Bayes is that the outcome of Hypothesis - h can be predicted given the evidence d.
- So, what are we trying to find… say as an example:
  - P(Saurabh is late| ML lecture)
  - We want to find out probability of Saurabh getting late if it is machine learning lecture. This is called as posterior probability
  - To compute this, we need:
    - P(Saurabh is late) : This is called as prior probability
    - P(ML lecture | Saurabh Is late) : probability that there is ML lecture when the Saurabh is late. Obtained from historical data
    - P(ML lecture) : probability of evidence-
    - ML lecture occurring.

- Lets talk more technically….

Naïve Bayesian Classification

- Simple Bayesian classifier works as follows:
  - Let D be training set.
  - It comprises of tuples,
  - Let each tuple be X=(x1,x2,x3…xn)
  - Assume there are m classes C1,c2,…Cm.

  - Given a tuple X, the classifier will predict that X belongs to class having highest posterior probability, conditioned on X.
  - So, the classifier predicts X to be of class Ci

  iff

$$P(Ci|X)>P(Cj|X) \text{ for } 1<=j<=m$$

- So, we need to maximize P(Ci|X).
- The class Ci for which P(Ci|X) is maximized is called maximum posterior hypothesis.
- So, by bayes therom we get
  - P(Ci|X)= (P(X|Ci)P(Ci))/P(X)
- Generally, P(X) is treated constant, hence only

  P(X|Ci)P(Ci) needs to be maximized.
- Now if class prior probabilities are not known then we assume that the probabilities are common. i.e.
  - P(C1)=P(C2)=....P(m)

Or else

P(Ci) = no. of training tuples of Ci /

D(no. of tuples in set)

- Hence, we need to maximize P(X|Ci), or else if the priorities were known we need to maximize P(X|Ci) and P(Ci).

- How do we get P(X|Ci) =

$$p(X \mid Ci) = \prod_{k=1}^{n} P(x_k \mid Ci)$$

- That is = P(x1|Ci)*P(x2|Ci)...*P(xn|Ci)

Play tennis example

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

What is hypothesis and evidence here?

- Hypothesis: Play (yes/ No)
- Evidence : Outlook = Sunny, temperature = hot and so on… are all evidences.
- What do we want to use Naïve Bayes for?
- Given some evidences, we want to predict if one can play tennis. That is we want to predict the class among yes/no.

Building the Naïve bayes model

- So, we have the train data available with us. We will build the model first.

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	Strong	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Weak	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Strong	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Weak	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

How are the values computed???

Building the Naïve bayes model

No. of times Outlook is Sunny , when Play = yes

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	Strong	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Weak	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Strong	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Weak	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

No. of times Outlook is Sunny /total no. of instances in training set where Play = yes

So, it is P(Outlook = Sunny| Play = yes) = 2/9.

What are the values in the Formula?

- P(h):-
  - P(plays = yes) = 9/14
  - P(plays = no) = 5/14

- P(outlook=sunny|yes) = 2/9
- P(outlook=overcast|yes)= 4/9
- ……..

- P(outlook=sunny|no)=3/5
- …..

Applying the model

- Given a new data, predict – play or no?
- Outlook=sunny
- Temperature = cool
- Humidity = High
- Windy = weak
- Play = ?

Using the rule

- P(Yes| E) =

  [P(Outlook=sunny|yes) *

  P(temperature=cool|yes) *

  P(humidity = high|yes)*

  P(windy=weak|yes)*

  P(plays=yes)] / P(E)

- P (E) is ignored (treated to be constant) as we need to find and compare values to other class.

Using the rule

- Thus the rule:
- P(Yes| E) =
  [P(Outlook=sunny|yes) *
  P(temperature=cool|yes) *
  P(humidity = high|yes)*
  P(windy=weak|yes)*
  P(plays=yes)]

=(2/9)*(3/9)*(3/9)*(3/9)*(9/14) = 0.0053
Similarly for P(No|E) = 0.0206.
Probability of not plays is higher,
hence the decision is not to play.

Some Queries…

- What will happen when value of any evidence in the built model becomes zero?
  - The entire result will be zero.
  - Generally technique called as smoothing is used.
- Will Naïve bayes guarantee to give correct output?
  - Even though the probability estimates for known evidences are low, yet the approach shows good classification results.
  - Many techniques like SVM, random forests have outperformed this approach, owing to the memory and speed factors, it is preferred.
- What is relation between Bayesian network and Naïve bayes?
  - At this instance, just remember that Naïve bayes are simple Bayesian networks that describes a particular class of Bayesian network.

Examples

Cough	Cold	Headache	Fever	Viral
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Severe	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Severe	Y	Y
N	Y	Severe	N	N
Y	Y	Mild	Y	Y

Cough	Cold	Headache	Fever	Viral
Y	N	Mild	N	?

Queries…

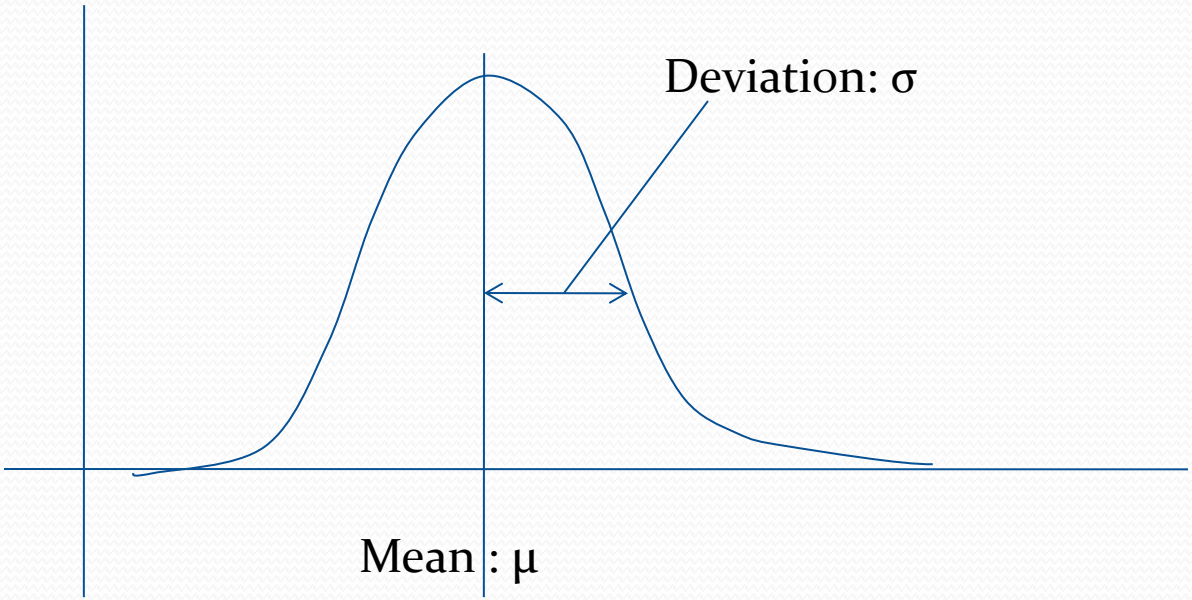
- This was about categorial attribute, what to do if the attribute is numeric, to be specific continuous?

Example

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	91	True	No

Gaussian distribution

- Central limit theorem : If there are more random un-co-related things happening, they follow Gaussian probability distribution.
- So, it looks like following:



Center/ target

Gaussian distribution

- Most values are near the middle.
- There are chances that you miss the target.
- Say for example:
  - I am looking for some ME student-
  1. Chances he is in the classroom on 5^{th} floor – center/target
  2. On the 6^{th} floor – near by
  3. On a ground floor lab – missed it...!!!

  : **Gaussian distribution**

- What if he is on ground floor only--- the output would be a parallel line :**Random distribution**

Why to use and when?

- It tells that any real observation will fall between any two real nos.
- Used in probabilistic calculations.
- The formula:

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

What is μ and σ?

- μ : Is the mean

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

- σ : Is the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n-1}}$$

What are we supposed to do now???

- Calculate
  - 1. μ and σ for : temp and humidity
    - How many values???
    - For play = yes
      - μ of temp and σ of temp
      - μ of humidity and σ of humidity
    - For play = no
      - μ of temp and σ of temp
      - μ of humidity and σ of humidity

Given unkown data- predict

- Outlook=sunny
- Temp = 61
- Humidity=70
- Windy = false

- For categorial attributes, you can calculate probabilities.
- For numeric values:
  - Use the Gaussian distribution so, with x=61 for temp and 70 for humidity, calculate the answer for play = yes
  - Use the Gaussian distribution so, with x=61 for temp and 70 for humidity, calculate the answer for play = no

Further calculations

- So, fill up the table with the calculations of the required data with respect to the unknown data and then calculate the final P(Play=yes|E) and P(Play=no|E). The higher one will be the answer.

References

- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.
- Tom M. Mitchell, (1997). Machine Learning, Singapore, McGraw- Hill.
- Quinlan, J.R. 1986. Induction of Decision trees. Machine Learning