COMP809 Data Mining and Machine Learning

LECTURER: DR AKBAR GHOBAKHLOU
SCHOOL OF ENGINEERING, COMPUTER AND MATHEMATICAL SCIENCES

Naïve Bayes Classification

Introduction and Agenda

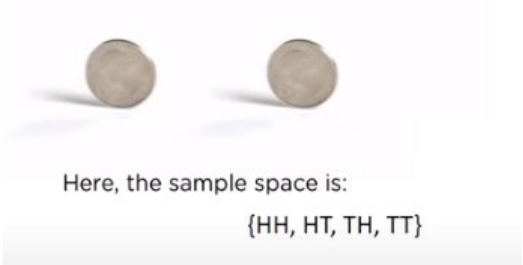
- What is Naive Bayes?
- Why do we need Naive Bayes?
- Understanding Naive Bayes Classifier
- Advantages of Naive Bayes Classifier

- A History of Bayes' Theorem

What is Naïve Bays?

- Thomas Bays introduced Bays Theorem in 1700s
- Uses the Bayes theorem for reasoning

Let's consider of the following example of tossing two coins



- The probability of getting two heads = 1/4
- The probability of at least one tail = 3/4
- The probability of the second coin being head given the first coin is tail = 1/2
- The probability of getting two heads given the first coin is a head = 1/2

Introducing Naïve Bays Theorem

- Each data feature contributes to a portfolio of evidence
- Assumes that all data features are statistically independent of each other – not always true

  - Probability of event *H* given evidence *E*:

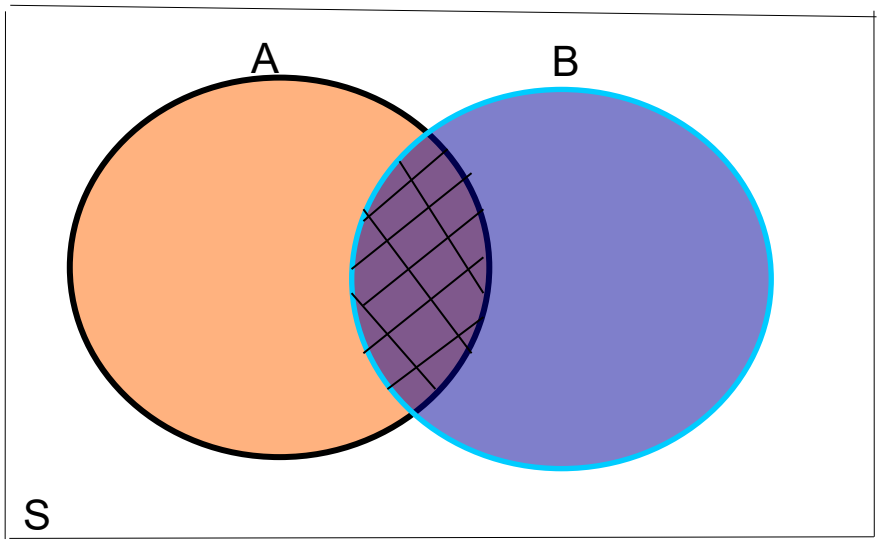
$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]}$$

  - *A priori* probability of *H*: $\Pr[H]$
    - Probability of event *before* evidence has been seen
  - *A posteriori* probability of *H*: $\Pr[H \mid E]$
    - Probability of event *after* evidence has been seen

Conditional Probability

- The conditional probability of an event *A* (given B) is the probability that an event *A* will occur given that another event, *B*, has occurred.

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



Useful Properties of Conditional Probabilities

Property 1. The Conditional Probability for Independent Events

If A and B are independent events, then:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Property 2. The Multiplication Rule for Conditional Probabilities

In an experiment involving two non-independent events A and B, the probability that both A and B occurs can be found in the following two ways:

$$P(B \cap A) = P(B)\,P(A|B)$$

or

$$P(A \cap B) = P(A)\,P(B|A)$$

Bayes Classifiers

Assumption: training set consists of instances of different classes described c_j as conjunctions of attributes values

Task: Classify a new instance d based on a tuple of attribute values into one of the classes $c_j \in C$

Key idea: assign the most probable class c_{MAP} using Bayes Theorem.

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)$$

MAP: Maximum A Posterior

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

Bayesian Rule

$$P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Example. 'Play Tennis' data

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*

Question: For the day <sunny, cool, high, strong>, what's the play prediction?

Naive Bayesian Classifier

- Given a training set, we can compute the probabilities

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5

Tempreature	Y	N
hot	2/9	2/5
mild	4/9	2/5

Windy	Y	N
true	3/9	3/5
false	6/9	2/5

$P(\text{Play}=Yes) = 9/14$ $P(\text{Play}=No) = 5/14$

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

Outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
Temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
Humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
Windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Play-tennis example: estimating P(x′|C)

Test Phase

- Given a new instance, predict its label

  x′=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

- Look up tables achieved in the learning phrase

P(Outlook=*Sunny*|Play=*Yes*) = 2/9
P(Temperature=*Cool*|Play=*Yes*) = 3/9
P(Huminity=*High*|Play=*Yes*) = 3/9
P(Wind=*Strong*|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=S*unny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*No*) = 5/14

$$\frac{\#\,days\ of\ \ playing\ tennise\ with\ strong\ wind}{\#\,days\ of\ \ playing\ tennise}$$

- Decision making with the MAP rule

P(*Yes*|x′) ≈ [P(*Sunny*|Yes)P(*Cool*|*Yes*)P(*High*|Yes)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053
P(*No*|x′) ≈ [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|x′) < P(*No*|x′), we label x′ to be "*No*".

Applying Naïve Bayes to Credit Scoring

		Counts	Counts	Probabilities	Probabilities
Independent Variable	Value	Good Risk	Poor Risk	Good Risk	Poor Risk
Debt	High	1	1	0.50	0.33
Debt	Low	1	2	0.50	0.67
Income	High	2	1	1.00	0.33
Income	Low	0	2	0	0.67
Married	Yes	2	2	1.00	0.67
Married	No	0	1	0	0.33
Total by Risk		2	3		

Naïve Bayes

- To classify an applicant, we make use of the previous table to work out Pr(Good Risk/Evidence) for that applicant
- In this example the hypothesis **H** can take one of two values: **Good Risk** or **Poor Risk**
- The evidence **E** is in three parts:
  - E1, which corresponds to Debt level
  - E2, corresponding to Income level
  - E3, corresponding to Marital Status
- From probability theory we have:

  Pr(E1, E2, E3)|H) = Pr(E1|H)*Pr(E2|H)*Pr(E3|H) under the assumption that E1, E2 and E3 are all independent of each other

Naïve Bayes

- We can now use Bayes theorem to work out the probability of the Good Risk and Poor Risk outcome for each individual
- Take Joe as an example: Joe has high Debt, has high income and is married.

$$\Pr(GR|Joe) = \frac{\Pr(D = H|GR) * \Pr(I = H|GR) * Pr(MS = Y|GR) * \Pr(GR)}{\Pr(E)}$$

$$= \frac{0.5*1.0*1.0*0.4}{\Pr(E)}$$

$$\Pr(PR|Joe) = \frac{\Pr(D = H|PR) * \Pr(I = H|PR) * Pr(MS = Y|PR) * \Pr(PR)}{\Pr(E)}$$

$$= \frac{0.33*0.33*0.67*0.6}{\Pr(E)}$$

- Since Pr(GR|Joe) > Pr(PR|Joe), Joe is classified as a good risk applicant

Naïve Bayes

- Likewise, classification for 4 further individuals appear in the table below

Name	Debt	Income	Married?	Risk Actual	Probability: Good Risk	Probability: Poor Risk	Risk Predicted
Joe	High	High	Yes	Good	0.82	0.18	Good
Sue	Low	High	Yes	Good	0.69	0.31	Good
John	Low	High	No	Poor	0	1.0	Poor
Mary	High	Low	Yes	Poor	0	1.0	Poor
Fred	Low	Low	Yes	Poor	0	1.0	Poor

Naïve Bayes – Laplace Correction

- A problem with Bayes theorem arises if no training samples appear for any class on a particular feature value

		Counts	Counts	Probabilities	Probabilities
Independent Variable	Value	Good Risk	Poor Risk	Good Risk	Poor Risk
Debt	High	1	1	0.50	0.67
Debt	Low	1	2	0.50	0.67
Income	High	2	3	1.00	1.0
Income	Low	0	0	0	0
Married	Yes	2	2	1.00	0.67
Married	No	0	1	0	0.33
Total by Risk		2	3		

Laplace Correction

- It would now not be possible to classify anyone with low income as both P(person|GR)= P(person|PR)=0.
- Thus we could not make any conclusions about persons whose income is low, *even though we have other evidence* for such people
- A simple solution exists for such situations
- $\Pr(X=x|C=c) = \frac{count(xc)+1}{count(c)+1}$ where c is the class value
- Provided count(c) is sufficiently large for all class values c, this does not distort probabilities significantly and thus solves the problem

Naive Bayes

- Models developed with Naive Bayes have good explanatory power, just as with Decision Trees
- In the credit scoring problem the Income and Marital Status attributes essentially determines the risk factor for a given individual (Why?)
- It is also true that the typical profile of a good risk applicant is someone who is married and has high income

Naïve Bayes

- Robust technique as it can deal with unknown/missing values
- Two types of scenarios for missing values:
  - Missing values in the training dataset

    Can be handled by not including these instances in the probability estimate calculations.
  - Missing values in the instance to be classified

    This needs care, if just one of the attribute values are missing, then the corresponding conditional probability can be set to 1 for ALL class outcomes and result will thus not be biased.

Naive Bayes for continuous data

- Continuous data can be handled by binning – done automatically by most DM products (such as Weka – see the supervised discretization option)
- An alternative scheme for continuous data is to compute probabilities using the Gaussian distribution
- For each class value, we assume that

$$\Pr(X = x|C = c) = N(\mu_c, \sigma_c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{(x-\mu_c)^2}{2\sigma_c^2}}$$

$\mu_c = \sum_{i=1}^{n}(x_i)_{|c_i=c}$ and $\sigma_c^2 = \frac{\sum_{i=1}^{n}(x_i-\mu_i)^2}{\#c-1}$ where n is the number of classes and #C is the number of data instances taking the value c

General Characteristics of Naïve Bayes

- One of the most efficient classification techniques as it makes only *one* pass through training data
- Naïve Bayes works well in practise even though features may not be statistically independent
- In cases where a large number of features are dependent on each other, accuracy could drop substantially and a more advanced version called Bayesian Network should be used

Bayesian Networks

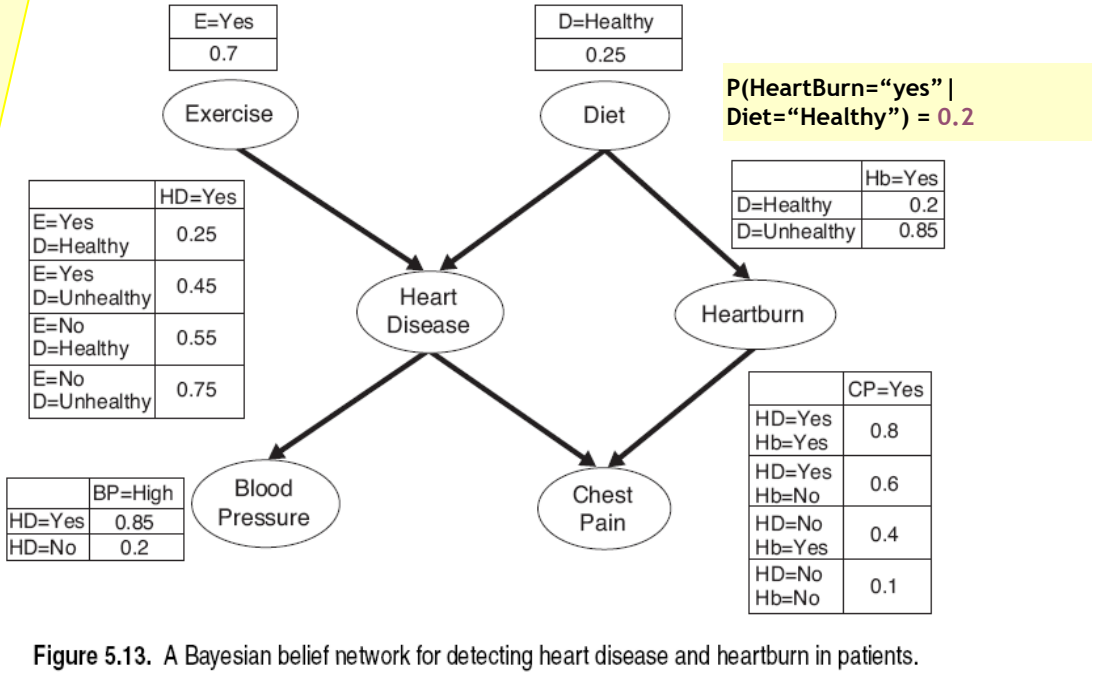


Figure 5.13. A Bayesian belief network for detecting heart disease and heartburn in patients.

P(HeartDisease="No"|Exercise="no", Diet ="Healthy") =
1 - P(HeartDisease="Yes"|Exercise="no", Diet ="Healthy") =1- 0.55 =
0.45

In general more accurate than Naïve Bayes but not as efficient as we need to learn structure of network – can be done by using an algorithm such as K2 (see reference at end)

References

- *Data Mining: Practical Machine Learning Tools and Techniques (3rd edition) /* Ian Witten, Eibe Frank; Elsevier, 2011, Chapter 4
  - Use of Kernel functions for estimating probabilities for Naïve Bayes:

    George H John and Pat Langley, Estimating Continuous Distributions in Bayesian Classifiers, available from Google scholar or AUT eLibrary:

  - K2 algorithm for learning structure of a Bayesian network from training data:

    G. Cooper and E. Herskovitz, A Bayesian method for the induction of probabilistic networks from data, Machine Learning, 9 (1992), 330–347.