Lecture 2: Bayesian Classification

Content

- Reminders from the last lecture
- Historical note
- Conditional probability
- Bayes theorem
- Bayesian classifiers
  - Example 1: Marketing promotions
  - Example 2: Items bought from Amazon
- Why use Bayesian classifiers?
- Lecture summary

Reminders from the last lecture

OR: $P(A \vee B) = P(B \vee A)$
$= P(A) + P(B)$

When A and B "Mutually exclusive" (i.e. they do not occur simultaneously)

AND: $P(A \wedge B)$

Independent A and B

$P(A) * P(B)$

Dependent A and B

Conditional rule:

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

P(B|A) means "the probability of observing event B *given that* event A has already been observed"

Ordering can be important:

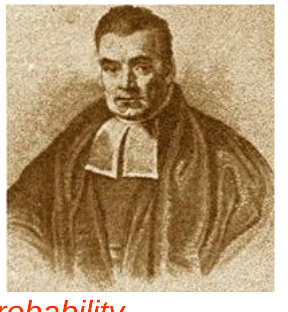
$P(A \vee B) = P(B \vee A)$

$P(A \wedge B) = P(B \wedge A)$

$P(A \mid B) \neq P(B \mid A)$

3

Historical note



- Thomas Bayes (b. 1702, London - d. 1761, Tunbridge Wells, Kent)
- Theologian and mathematician
- Established a mathematical basis for *probability inference* (a means of calculating, from the frequency with which an event has occurred in prior trials, the probability that it will occur in future trials).
- Essay "Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the *Philosophical Transactions of the Royal Society*.

Source: The Encyclopaedia Britannica.
http://www.britannica.com/EBchecked/topic/56807/Thomas-Bayes

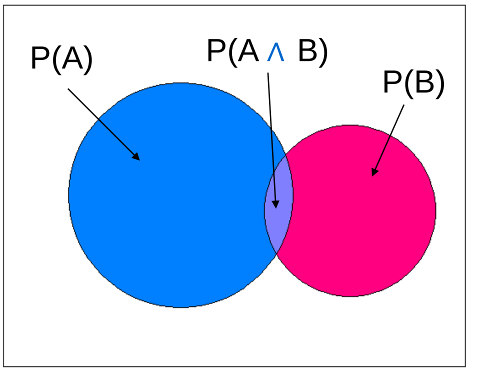
4

Conditional probability

Venn diagram.

- Area of the rectangle is 1
- Area of each region gives the probability of the event(s) associated with that region

P(A|B) means "the probability of observing event A *given that* event B has already been observed"



P(A|B): how much of the time that we see B do we also see A? (i.e. the ratio of the purple region to the pink region)

$P(A|B) = P(B \wedge A) / P(B)$ Conditional Probability

Getting to Bayes' theorem from conditional probabilities

- To derive the theorem, we start from the definition of **conditional probability**. The probability of event *A* given event *B* is
  $P(A|B) = P(B \wedge A) / P(B)$
- Likewise, the probability of event *B* given event *A* is
  $P(B|A) = P(A \wedge B) / P(A)$
- Rearranging and combining these two equations, we find
  $P(A|B)\, P(B) = P(B \wedge A) = P(B|A)\, P(A)$
- Dividing both sides by *P*(*B*), providing that it is non-zero, we obtain **Bayes' theorem**:

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

What is a Bayesian classifier?

- Bayesian Classifiers are **statistical** classifiers
  - based on **Bayes' Theorem**
- They can predict the probability that a particular sample is a member of a particular class
- The simplest Bayesian Classifier is known as the **Naïve Bayesian Classifier** based on an independence assumption
- We assume that values given for one variable are not influenced by values given to another variable. No relationship exists between them
- Although the independence assumption is often a bold assumption to make, performance is still often comparable to Decision Trees and Neural Network classifiers

Example 1: Marketing promotions

Data related to marketing promotions and gender of the customer

Mag. Promotion	TV Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Y	N	N	N	M
Y	Y	Y	N	F
N	N	N	N	M
Y	Y	Y	Y	M
Y	N	Y	N	F
N	N	N	N	F
Y	N	Y	Y	M
N	Y	N	N	M
Y	N	N	N	M
Y	Y	Y	Y	F

Let's consider sex/gender as the output attribute whose value is to be predicted

Example 1: Marketing promotions

Examples of things we can derive from our dataset:

- 4 males took advantage of the Mag. Promo; they represent 2/3 of the total male population
- 3/4's of females purchased the Mag. Promo

We want to classify a new instance (or customer), called Lee

We are told the following holds true for our new customer

Mag. Promo = Y
TV Promo = Y
LI Promo = N
C.C. Ins. = N

This is our evidence E

2 hypothesis:
H1: Lee is male
H2: Lee is female

We want to know if Lee is female or male

We apply Bayes' classifier and compute a probability for each hypothesis

Example 1: Marketing promotions

Using a distribution Table

- List the distribution of the output attribute value (Sex) for each input attribute (Mag. Promo, TV Promo, LI Promo, C.C. Ins)

	Mag Promo		TV Promo		LI Promo		C.C. Ins	
Sex	M	F	M	F	M	F	M	F
Y	4	3	2	2	2	3	2	1
N	2	1	4	2	4	1	4	3
Ratio: Yes/Total	4/6	3/4	2/6	2/4	2/6	3/4	2/6	1/4
Ratio: No/Total	2/6	1/4	4/6	2/4	4/6	1/4	4/6	3/4

So for example,
4 males answered Y to the Mag Promo

Check ratio for Y/T and N/T add to 1 for each column

2 out of the total of 6 males answered
Y to the LI Promo

Example 1: Hypothesis H1 and H2

H1: Lee is male

$$P(sex = M \mid E) = \frac{P(E \mid sex = M)\, P(sex = M)}{P(E)}$$

Bayes' Theorem

Starting with P(E | sex = M) … This is

P(Mag. Promo = Y, TV Promo = Y, LI Promo = N, C.C. Ins = N | sex = M)

We have (mathematical justification):

$P(E1 \wedge E2 \wedge E3 \wedge E4 \mid M)\, P(M) =$

$P(E1 \wedge E2 \wedge E3 \wedge E4 \wedge M) =$

$P(E1 \mid E2 \wedge E3 \wedge E4 \wedge M)\, P(E2 \wedge E3 \wedge E4 \wedge M) =$

$P(E1 \mid M)\, P(E2 \wedge E3 \wedge E4 \wedge M) = \ldots$

$P(E1 \mid M)\, P(E2 \mid M)\, P(E3 \mid M)\, P(E4 \mid M)\, P(M)$

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

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

* Assumption**: E1, … E4 are conditionally independent given M**, i.e.. the information added by knowing that E2, … E4 have happened does not add much to P(E1 | M) and is forgotten. This is not always correct, it is an approximation, but often works well (and fast!).

Example 1: H1 and H2

H1: Lee is male

$$P(sex = M \mid E) = \frac{P(E \mid sex = M)\, P(sex = M)}{P(E)}$$

Bayes' Theorem

Starting with P(E| sex = M)

This is calculated by multiplying the conditional probability values for each piece of evidence. So, overall conditional probability is the product of the following:

P(Mag. Promo = Y | sex = M) = 4/6
P(TV Promo = Y | sex = M) = 2/6
P(LI Promo = N | sex = M) = 4/6
P(C.C. Ins = N | sex = M) = 4/6

(These values are dictated by our current case.. See earlier slide – distribution table)

Example 1: H1 and H2

Therefore, the conditional probability that sex = M is:

P(E | sex = M) = (4/6) * (2/6) * (4/6) * (4/6) = 8/81

Part 1 done, now let's look at computing P(sex = M)

This is the probability of a male customer (with no knowledge of any other information)

In this case, it's just what fraction of our total population is male… 6/10 or 3/5

So we now have:

$$P(\text{sex} = M \mid E) = \frac{(8/81) * (3/5)}{P(E)}$$

Let's now to the female of the species…

Example 1: H1 and H2

H2: Customer is female

$$P(\text{sex} = F \mid E) = \frac{P(E \mid \text{sex} = F)\, P(\text{sex} = F)}{P(E)}$$

Bayes' Theorem

Starting with P(E| sex = F)

This is calculated by multiplying the conditional probability values for each piece of evidence. So, overall conditional probability is the product of the following:

P(Mag. Promo = Y | sex = F) = 3/4
P(TV Promo = Y | sex = F) = 2/4
P(LI Promo = N | sex = F) = 1/4
P(C.C. Ins = N | sex = F) = 3/4

Example 1: H1 and H2

Therefore, the conditional probability that sex = F is:

P(E | sex = F) = (3/4) * (2/4) * (1/4) * (3/4) = 9/128

Part 1 done for the females, now let's look at computing P(sex = F)

This is the probability of a female customer (with no knowledge of any other information)

In this case, it's just what fraction of our total population is female… 4/10 or 2/5

So we now have:

$$P(\text{sex} = F \mid E) = \frac{(9/128) * (2/5)}{P(E)}$$

15

Example 1: H1 and H2

So we have:

H1:

$$P(\text{sex} = M \mid E) = \frac{(8/81) * (3/5)}{P(E)}$$

H2:

$$P(\text{sex} = F \mid E) = \frac{(9/128) * (2/5)}{P(E)}$$

P(E) = probability of the evidence having occurred. The value is the same for both components.

So all we need to know is which is has the greater probability, H1 or H2?

(8/81)*(3/5) = 8/135 ~0.06 (9/128)*(2/5) = 9/320 ~ 0.03

Hence, Bayes' classifier tells us that Lee is most likely a male.

Calculating P(E), i.e. the (conditionally independent) probabilities of Mag. Promo, TV Promo, not LI Promo and not CC Promo

P(E) = (7/10) * (4/10) * (5/10) * (7/10) = 0.098

we have: P(sex = F | E) = 0.2815 < 0.5926 = P(sex = M | E)

16

Example 2: Items bought from Amazon

CDs	Books	DVDs	Videos	Region
Y	Y	N	N	Stirling
Y	N	Y	N	Glasgow
Y	N	Y	Y	Glasgow
Y	Y	Y	N	Glasgow
N	N	Y	N	Stirling
N	Y	Y	Y	Stirling
Y	N	Y	N	Stirling
Y	Y	Y	Y	Glasgow

1. What proportion of Glasgow customers buy books?
2. What proportion of all customers buy DVDs?
3. Given a new customer that we knows buys Videos, is it more likely that they live in Glasgow or Stirling?

Example 2: Items bought from Amazon

	CDs G	CDs S	Books G	Books S	DVDs G	DVDs S	Videos G	Videos S
Y	4	2	2	2	4	3	2	1
N	0	2	2	2	0	1	2	3
Ratio Y/Tot	1	½	½	½	1	¾	½	¼
Ratio N/Tot	0	½	½	½	0	¼	½	¾

Example 2: Items bought from Amazon

$$P(\text{Glasgow} \mid \text{videos}) = \frac{P(\text{videos} \mid \text{Glasgow})\, P(\text{Glasgow})}{P(\text{videos})}$$

P(videos | Glasgow) = ½

P(Glasgow) = ½

P(Glasgow | videos) = (½ * ½)/(3/8) = 2/3

$$P(\text{Stirling} \mid \text{videos}) = \frac{P(\text{videos} \mid \text{Stirling})\, P(\text{Stirling})}{P(\text{videos})}$$

P(videos | Stirling) = ¼

P(Stirling) = ½

P(Stirling | videos) = (¼ * ½)/(3/8) = 1/3

More likely to be from Glasgow

Why use Bayesian classifiers?

- No classification method has been found to be superior over all others in every case (i.e. a data set drawn from a particular domain of interest)
- Methods can be compared based on:
  - accuracy
  - interpretability of the results
  - robustness of the method with different datasets
  - training time
  - scalability

Lecture summary

- Bayesian classification can help to predict information we do not know using information we do know and the likelihood of certain patterns in the data occurring.

- Such approaches can be useful in marketing products.

- We can learn information about customers such that we can predict how likely a customer may be to:
    - Be interested in a new product we’re offering
    - Change their loyalty and start shopping elsewhere
    - Purchase items together