Probabilistic and Bayesian Analytics

Chenghua Lin
Chenghua.Lin@abdn.ac.uk

Road Map

- Probabilistic and Bayesian Analytics
- Classification
  - Naïve Bayes Classifier
  - Support vector machines (SVM)
  - Decision Trees
- Association Rule Mining
- Feature Engineering
- Data visualization
- Case study
- Data Mining Issues

Today's Lecture

- Probabilistic and Bayesian Analytics
- Classification
  - Naïve Bayes Classifier
  - Support vector machines (SVM)
  - Decision Trees
- Association Rule Mining
- Feature Selection
- Visualization I and II
- Case study
- Data Mining Issues

Probability

- The world is a very uncertain place
- Probability: a mathematical framework for reasoning about uncertainty
  - How likely next Monday is going to rain?
  - You have a headache. What's the chance you have got flu?
  - Many other similar examples …

What we’re going to do

- We will review the fundamentals of probability.
- It’s really going to be worth it
- We will keep mathematics to the minimum

Discrete Random Variables

- A is a **Boolean-valued random variable** if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- Examples
  - A = The UK PM in 2023 will be female
  - A = You wake up tomorrow with a headache
  - A = You have Ebola…

The Axioms of Probability

- $0 <= P(A) <= 1$

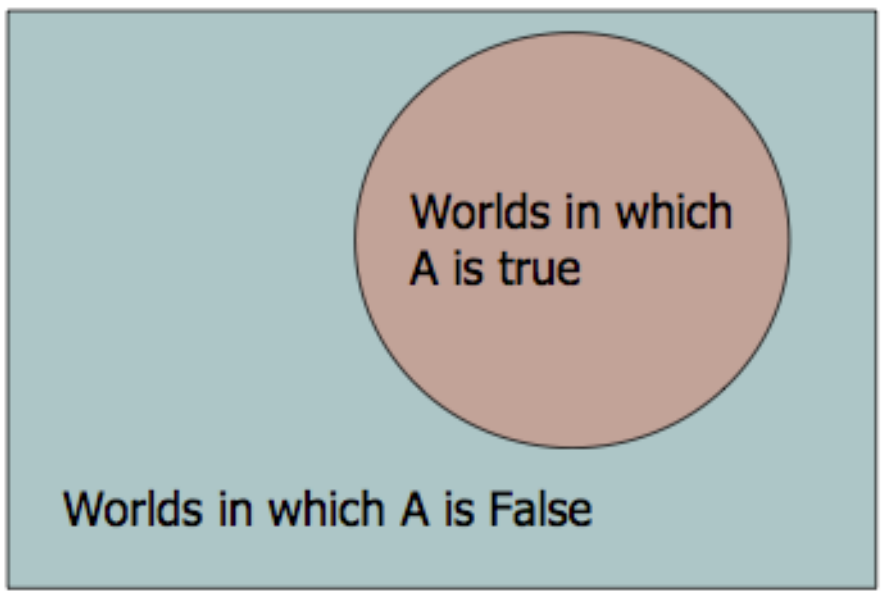
- $P(\Omega) = 1$

Probabilities

- We write P(A) as "**the fraction of possible worlds in which A is true**"

Sample space of all possible worlds **(Ω)**

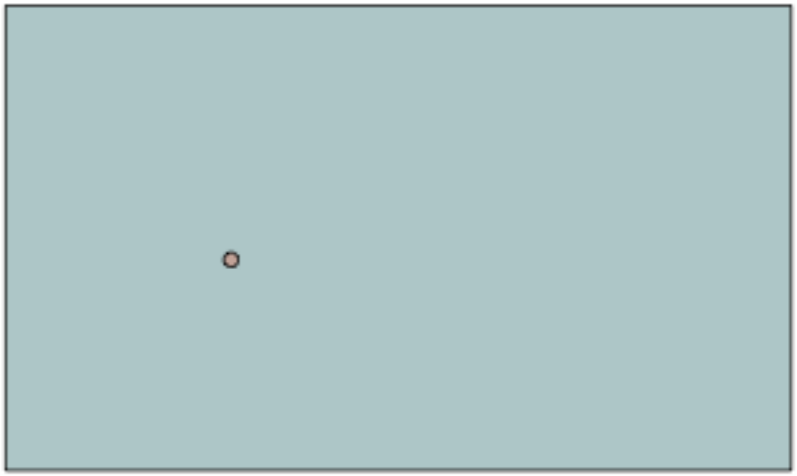
Its area is 1



P(A) = Area of reddish oval

Interpreting the axioms

- 0<=P(A)<=1
- P(Ω) = 1

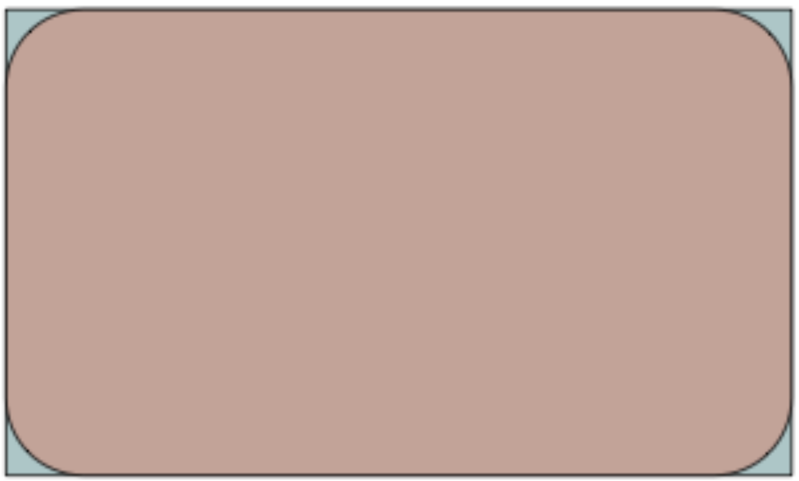


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

- 0<=P(A)<=1
- P(Ω) = 1



The area of A can’t get any bigger than 1

And an area of 1 would mean all worlds will have A true

The Axioms of Probability

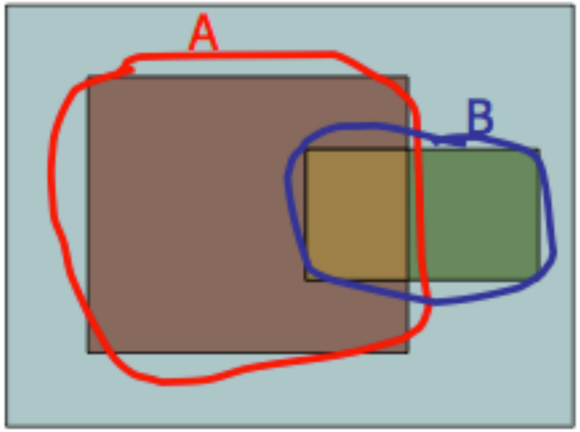
- $0 <= P(A) <= 1$

- $P(\Omega) = 1$

- $P(A \text{ or } B)=P(A)+P(B)-P(A \text{ and } B)$

Interpreting the axioms

- $0 <= P(A) <= 1$
- $P(\Omega) = 1$
- P(A or B)=P(A)+P(B)-P(A and B)



Multivalued Random Variables

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of {v1,v2, .. vk}
  - i.e., the elements of the value space has to be **mutually exclusive**
- Thus... $P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$

$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

An easy fact about Multivalued Random Variables

- Using the axioms of probability
  - 0<=P(A)<=1
  - P(Ω) = 1
  - P(A or B)=P(A)+P(B)-P(A and B)
- And assuming that A obeys

$$P(A=v_i \wedge A=v_j)=0 \text{ if } i \neq j$$
$$P(A=v_1 \vee A=v_2 \vee A=v_k)=1$$

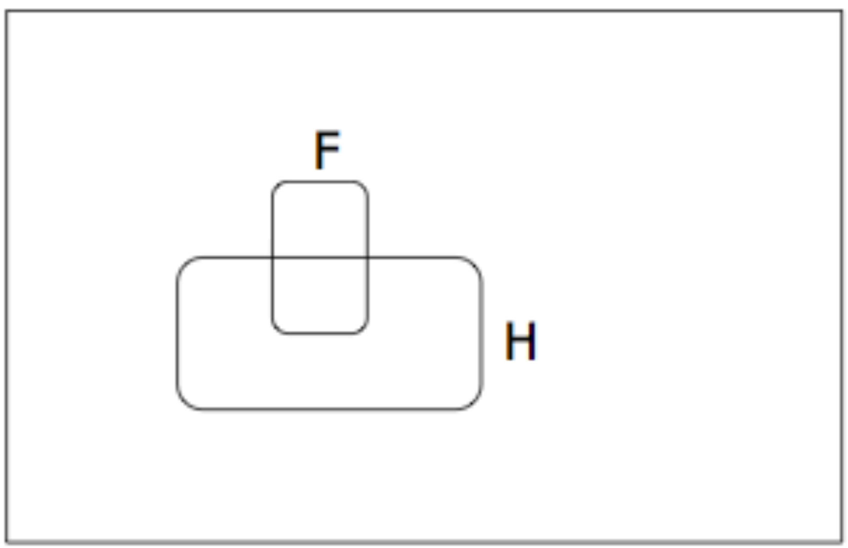
- It is easy to prove that

$$P(A=v_1 \vee A=v_2 \vee A=v_i)=\sum_{j=1}^{i} P(A=v_j)$$

- And thus we can prove $\displaystyle\sum_{j=1}^{k} P(A=v_j)=1$

Conditional Probability

- P(A|B) = The probability of an event (A), given that another (B) has already occurred.

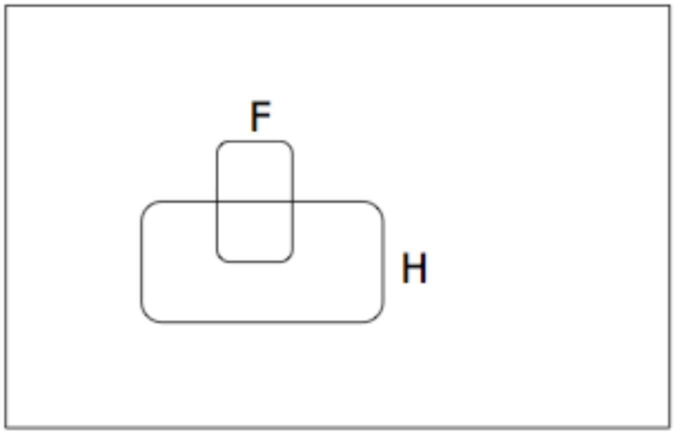


H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Conditional Probability



P(H|F) = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\#\text{worlds with flu and headache}}{\#\text{worlds with flu}}$$

$$= \frac{\text{Area of “H and F” region}}{\text{Area of “F” region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

H = “Have a headache”
F = “Coming down with Flu”
P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

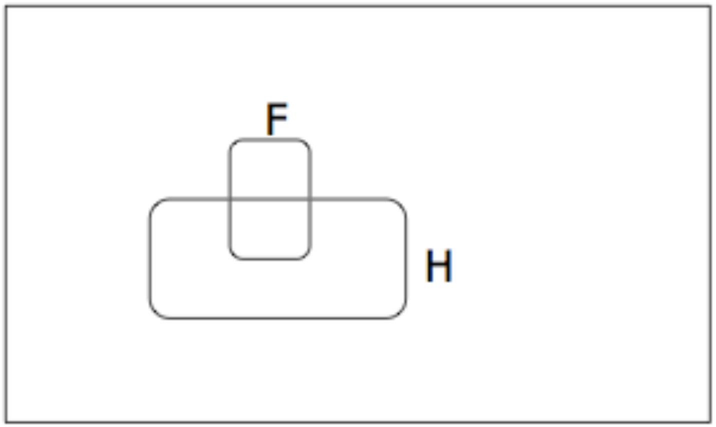
Definition of Conditional Probability

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

- Corollary: The Chain Rule

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

Probabilistic Inference



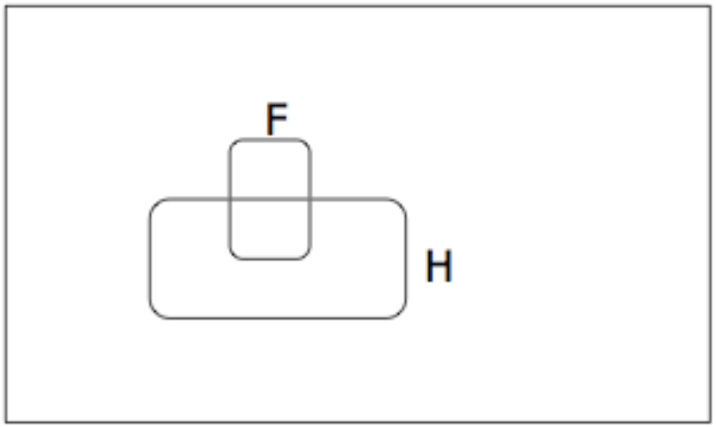
H = "Have a headache"
F = "Coming down with Flu"

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

Probabilistic Inference



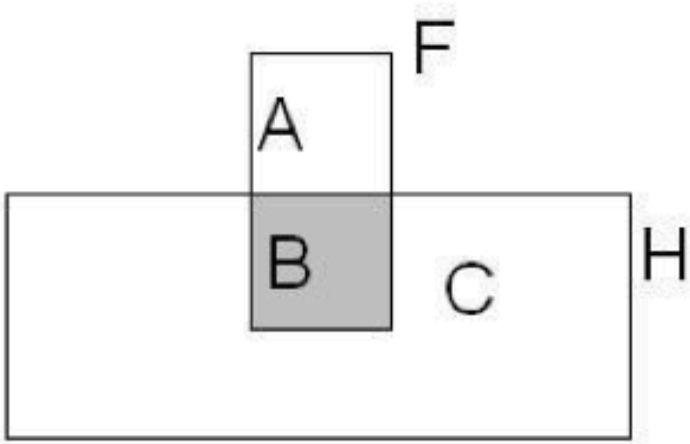
H = “Have a headache”
F = “Coming down with Flu”

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

$P(F \wedge H) = \ldots$

$P(F|H) = \ldots$

Another way to understand the intuition



Let's say we have P(F), P(H), and P(H|F), like in the example in class.

Areawise, $P(F) = A + B$, $P(H) = B + C$,

Also, $P(H|F) = \dfrac{B}{A + B}$

Thus, to get the opposite conditional probability, ie, P(F|H), we need to figure out $\dfrac{B}{B + C}$

Since we know B / (A+B), we can get B / (B+C) by multiplying by (A+B) and dividing by (B+C). But since we already calculated, A+B = P(F), and B+C = P(H), so we are actually multiplying by P(F) and dividing by P(H). Which is Bayes Rule:

$$P(F|H) = P(H|F) * \frac{P(F)}{P(H)}$$

What we just did...

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

This is the FAMOUS **Bayes Rule**

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

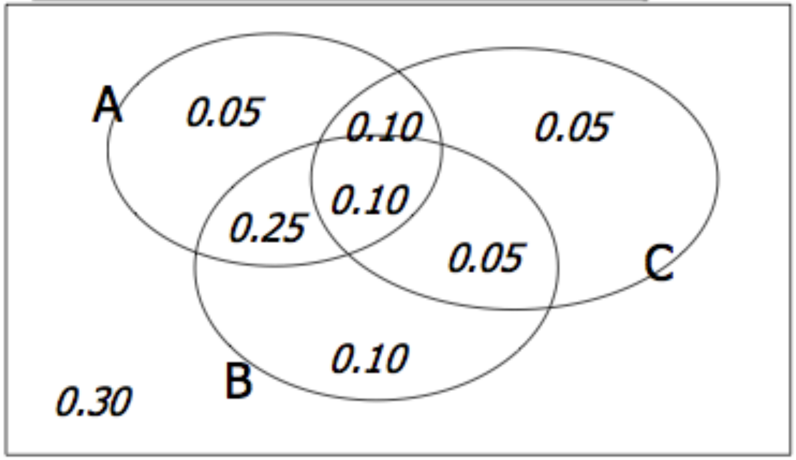
The Joint Distribution

Joint distribution: the probability of two or more events occurring together.

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

- Once you have the joint distribution you can ask for the probability of any logical expression involving your attribute

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

P(Poor Male) = 0.4654

P(Poor) = 0.7604

P(Male | Poor)
= 0.4654 / 0.7604
= 0.612

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

What you should know

- The Axioms of Probability
- Conditional probability
- Joint probability
- Bayes rule

Acknowledgement

- Some part of the lecture notes are from Andrew W. Moore’s tutorial.