Probabilistic and Bayesian Analytics

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Significantly modified by Lyle Ungar

What we're going to do

- **We will review the fundamentals of probability.**
- **It's really going to be worth it**
  - Much of this course builds on probabilities
  - E.g. Naïve Bayes, Logistic regression, Bayes nets, LDA …

Key concepts

- **Sample spaces, Events and Random Variables**
- **Expectation**
- **Probability distributions**
  - Discrete, continuous and joint distributions
  - Marginalization
  - PDFs and CDFs
- **Rules of probability**
  - Conditional probability, Bayes rule, Chain rule
  - Independence, Conditional independence

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 US president in 2023 will be male**
- **A = You wake up tomorrow with a headache**
- **A = You have Ebola**

How is this slide wrong?

A Random variable maps from an element of the sample space to a real number

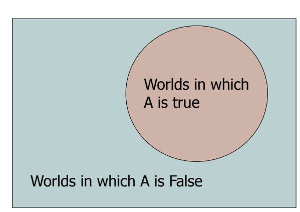
Probabilities

- **We write P(A) as "the fraction of possible worlds in which A is true"**
- **We could at this point spend 2 hours on the philosophy of this.**
- **But we won't.**

Visualizing A

Event space of all possible worlds

Its area is 1



P(A) = Area of reddish oval

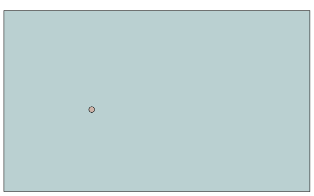
The Axioms of Probability

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

Where do these axioms come from? Were they "discovered"? Answers coming up later.

Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)



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

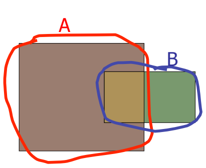
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

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

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

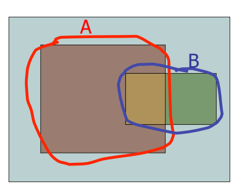
Interpreting the axioms

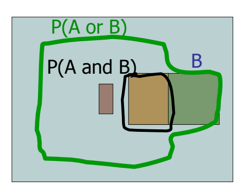
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)



Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

These Axioms are Not to be Trifled With

- **There have been attempts to do different methodologies for uncertainty**
  - Fuzzy Logic
  - Three-valued logic
  - Dempster-Shafer
  - Non-monotonic reasoning
- **But the axioms of probability are the only system with this property:**

If you gamble using them you can't be unfairly exploited by an opponent using some other system [di Finetti 1931]

Theorems from the Axioms

- **0 <= P(A) <= 1,**
- **P(True) = 1,**
- **P(False) = 0**
- **P(A or B) = P(A) + P(B) - P(A and B)**

From these we can prove:

P(not A) = P(~A) = 1-P(A)

Side Note

- **I am inflicting these proofs on you for two reasons:**
  1. These kind of manipulations will need to be second nature to you if you use probabilistic analytics in depth
  2. Suffering is good for you

Another important theorem

- **0 <= P(A) <= 1,**
- **P(True) = 1,**
- **P(False) = 0**
- **P(A or B) = P(A) + P(B) - P(A and B)**

From these we can prove:

P(A) = P(A ^ B) + P(A ^ ~B)

Random Variables

- **A *random variable* maps from an element of a sample space to a discrete or real property of that element**
- **Examples**
  - S(x): person x is male → 1
    x is female → 0
  - A(x): = person x → x's age

Technically: random variable maps from an element of the sample space to a real number

We assign a probability to each outcome

- P(S(x) = 1)
- P(A(x) = 25)

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 $\{v_1, v_2, .. v_k\}$**
- **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(True) = 1, P(False) = 0
  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's easy to prove that**

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

An easy fact about Multivalued Random Variables:

- **Using the axioms of probability…**
  0 <= P(A) <= 1, P(True) = 1, P(False) = 0
  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's 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

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

Another fact about Multivalued Random Variables:

- **Using the axioms of probability…**
  0 <= P(A) <= 1, P(True) = 1, P(False) = 0
  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's easy to prove that

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

Another fact about Multivalued Random Variables:

- **Using the axioms of probability…**
  0 <= P(A) <= 1, P(True) = 1, P(False) = 0
  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' s easy to prove that

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

- And thus we can prove

$$P(B) = \sum_{j=1}^{k} P(B \wedge A = v_j)$$

Elementary Probability in Pictures

- **P(~A) + P(A) = 1**

Elementary Probability in Pictures

- **P(B) = P(B ^ A) + P(B ^ ~A)**

Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

Elementary Probability in Pictures

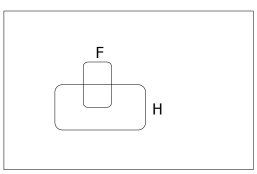
$$P(B) = \sum_{j=1}^{k} P(B \wedge A = v_j)$$

Conditional Probability

- **P(A|B) = Fraction of worlds in which B is true that also have A true**

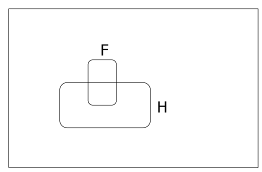
H = "Have a headache"
F = "Coming down with Flu"

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



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

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

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

= #worlds with flu and headache

#worlds with flu

= Area of "H and F" region

Area of "F" region

= P(H ^ F)

P(F)

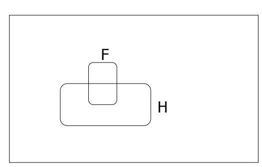
Definition of Conditional Probability

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

Corollary: The Chain Rule

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

Probabilistic Inference



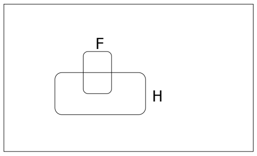
H = "Have a headache"
F = "Coming down with Flu"

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"

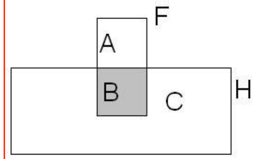
P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

P(F ^ H) = ...

P(F|H) = ...

Another way to understand the intuition

Thanks to Jahanzeb Sherwani for contributing this explanation:



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) = $\frac{B}{A + B}$

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

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…

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

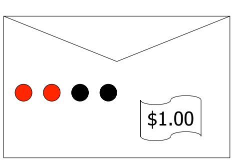
This is Bayes Rule

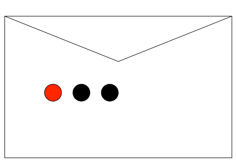
Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Bayes rule made easy

- **B** **~ B**
- **A** 0.01 0.1
- **~A** 0.09 0.8
- **Counts instead of probabilities**
- **B** **~ B**
- **A** 1 10
- **~A** 9 80

Using Bayes Rule to Gamble



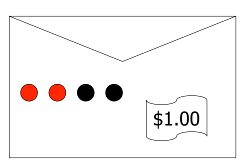


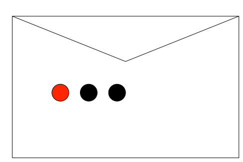
The "Win" envelope has a dollar and four beads in it

The "Lose" envelope has three beads and no money

Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay?

Using Bayes Rule to Gamble





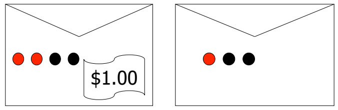
The "Win" envelope has a dollar and four beads in it

The "Lose" envelope has three beads and no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?
Suppose it's red: How much should you pay?

Calculation…



More General Forms of Bayes Rule

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

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

More General Forms of Bayes Rule

$$P(A = v_i|B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n_A} P(B|A = v_k)P(A = v_k)}$$

Useful Easy-to-prove facts

$$P(A|B) + P(\neg A|B) = 1$$

$$\sum_{k=1}^{n_A} P(A = v_k|B) = 1$$

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

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

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

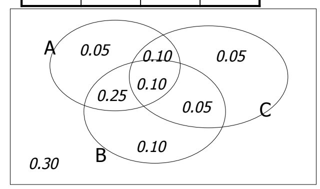
The Joint Distribution

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

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M 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.



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

One you have the JD you can ask for the probability of any logical expression involving your attribute

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

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

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

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) = 0.7604

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

Inference with 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

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

Inference with 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

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

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

Inference is a big deal

- **I've got this evidence. What's the chance that this conclusion is true?**
  - I've got a sore neck: how likely am I to have meningitis?
  - I see my lights are out and it's 9pm. What's the chance my spouse is already asleep?
- **There's a thriving set of industries growing based around Bayesian Inference, including**
  - Medicine, Pharma, Help Desk Support, Engine Fault Diagnosis

Where do Joint Distributions come from?

- **Idea One: Expert Humans**
- **Idea Two: Simpler probabilistic facts and some algebra**

Example: Suppose you knew

P(A) = 0.7
P(B|A) = 0.2
P(B|~A) = 0.1

P(C|A^B) = 0.1
P(C|A^~B) = 0.8
P(C|~A^B) = 0.3
P(C|~A^~B) = 0.1

Then you can automatically compute the JD using the chain rule

P(A=x ^ B=y ^ C=z) =
P(C=z|A=x^ B=y) P(B=y|A=x) P(A=x)

In another lecture: Bayes Nets, a systematic way to do this.

Where do Joint Distributions come from?

- **Idea Three: Learn them from data!**

Prepare to see one of the most impressive learning algorithms you'll come across in the entire course….

Learning a joint distribution

Build a JD table for your attributes in which the probabilities are unspecified

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

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

Fraction of all records in which A and B are True but C is False

Example of Learning a Joint

- This Joint was obtained by learning from three attributes in the UCI "Adult" Census Database [Kohavi 1995]

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

Key concepts

- **Sample spaces, Events and Random Variables**
- **Expectation**
- **Probability distributions**
  - Discrete, continuous and joint distributions
  - Marginalization
  - PDFs and CDFs
- **Rules of probability**
  - Conditional probability, Bayes rule, Chain rule
  - Independence, Conditional independence