Probability continued

Project Pitches – This Thursday in Lab

- 1 minute
- Worth 1% of your mark
- Convince your classmates it's a good project idea
- Include
  - what is the task?
  - why is it an important problem?
  - how will you solve it? is there data already available?

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)

Monotonicity: if A is a subset of B, then P(A) <= P(B)

Proof:
- A subset of B ➔ B = A + C for C=B-A
- A and C are disjoint ➔ P(B) = P(A or C)=P(A) + P(C)
- P(C) >= 0

- So P(B) >= P(A)

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)

Theorem: P(~A) = 1 - P(A)

Proof:
- P(A or ~A) = P(True) = 1
- A and ~A are disjoint ➔ P(A or ~A)=P(A) + P(~A)
  ➔ P(A) + P(~A) = 1

  ….then solve for P(~A)

Multivalued Discrete 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 $\{v_1, v_2, \ldots v_k\}$
- *Example:* A=*{1,2,3….,20}:* good for 20-sided dice games
- Notation: let's write the event AHasValueOf*v* as "A=*v*"

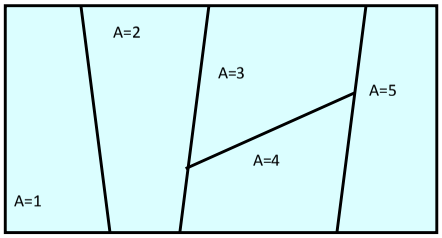
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

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

Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$ (Law of total probability)

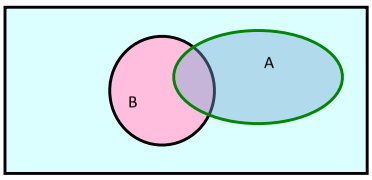


Definition of Conditional Probability

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

We say "probability of A given b"

Foundation for Bayes' Rule!



Definition of Conditional Probability

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

Corollary: The Chain Rule

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

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

Independent Events

- Definition: two events A and B are *independent* if P(A and B)=P(A)*P(B)
- Intuition: knowing A tells us nothing about the value of B (and vice versa)
- From chain rule

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

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

- You frequently need to assume the independence of *something* to solve a learning problem.

Continuous Random Variables

- The discrete case: sum over all values of A is 1

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

- The continuous case: infinitely many values for A and the *integral* is 1

$$\int_{-\infty}^{\infty} f_P(x)\,dx = 1$$

f(x) is a probability density function (pdf)

1. 0<=P(A) <= 1
2. Pr(True) = 1
3. P(A or B) = P(A) + P(B) - P(A and B)

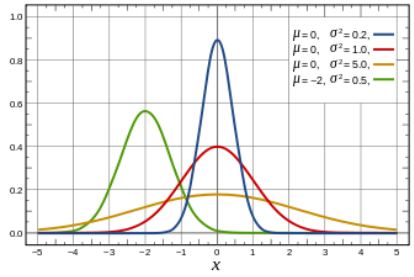
also....

$$\forall x,\ f_P(x) \geq 0$$

Continuous Random Variables

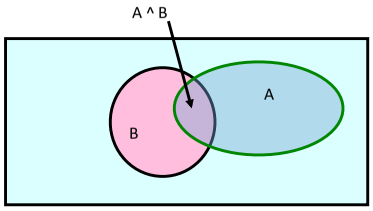
Gaussian probability density with parameters
- mean μ
- standard deviation σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Bayes Rule

- let's write two expressions for P(A ^ B)



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

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

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**

...by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter.... necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*...

Other Forms of Bayes Rule

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

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

Applying Bayes Rule

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

A = you have the flu, B = you just coughed

Assume:
P(A) = 0.05

Also assume the following information is known to you
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

Joint distribution

- Probability of >1 thing happening at the same time
  - Probability it will rain today and I forgot my umbrella
    - P(rain=true,umbrella=false)

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

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.

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

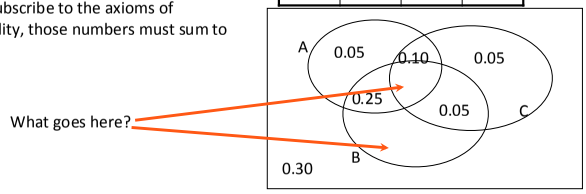
The Joint Distribution

Recipe for making a joint distribution of M variables:

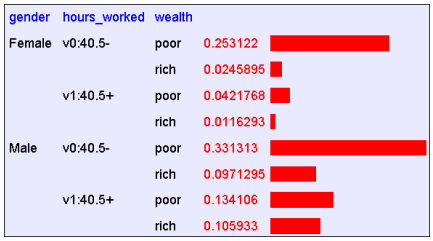
1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

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

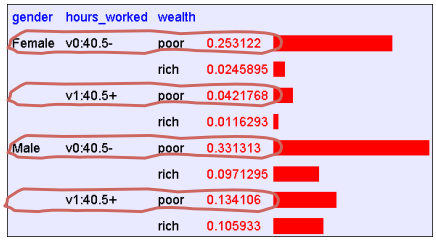


Joint Probability Distribution



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

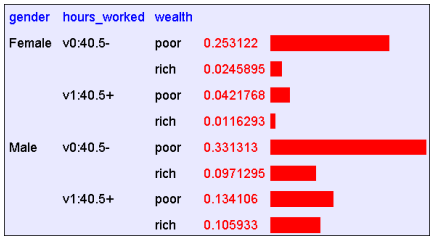
Using the Joint Distribution



P(Poor) = 0.7604

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

Inference with the Joint



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

Maximum Likelihood Estimation (MLE)

Rich vs Poor

What is the probability of a person being rich, given you know nothing else about that person?

3:2

Why 3/5?

We assume that the wealth of the people in our dataset *D* is independently distributed

θ= Probability of being rich = P(rich)

? = Probability of being poor = P(poor)

D = { r, p, r, r, p} α_r = # rich α_p = # poor

$$P(D) = P(r \text{ and } p \text{ and } r \text{ and } r \text{ and } p)$$

A little math

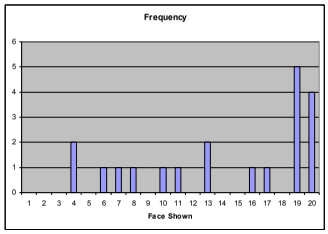
$$\underset{\theta}{\text{argmax}}\ P(D) = (1-\theta)^{\alpha_F} * \theta^{\alpha_H}$$

That's Maximum Likelihood Estimation (MLE)

It's not always the best solution…

Issues with MLE estimate

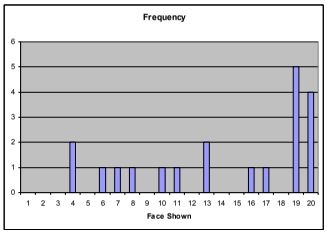
I bought a loaded 20-faced die (d20) on EBay…but it didn't come with any specs. How can I find out how it behaves?



1. Collect some data (20 rolls)
2. Estimate P(i)=CountOf(rolls of i)/CountOf(any roll)

Issues with MLE estimate

I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?

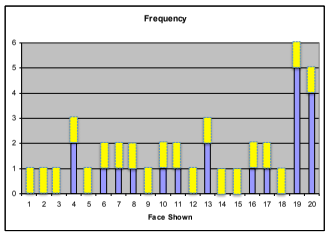


P(1)=0
P(2)=0
P(3)=0
P(4)=0.1
…
P(19)=0.25
P(20)=0.2

But: Do I really think it's *impossible* to roll a 1,2 or 3?

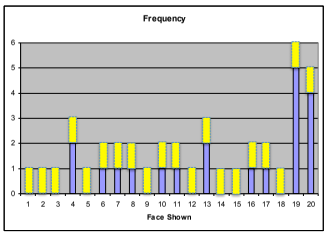
A better solution

I bought a loaded d20 on EBay…but it didn't come with any specs.
How can I find out how it behaves?



0. *Imagine* some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate P(i)

A better solution?



P(1)=1/40
P(2)=1/40
P(3)=1/40
P(4)=(2+1)/40
…
P(19)=(5+1)/40
P(20)=(4+1)/40=1/8

MAP = maximum a posteriori estimate

$$\hat{P}(i) = \frac{CountOf(i)+1}{CountOf(ANY)+CountOf(IMAGINED)}$$

0.2 *vs.* 0.125 – really different! Maybe I should "imagine" less data?

What if we know that poor people are much more common than rich people?

We have a belief about θ

- $P(\theta|D) = P(D|\theta)*P(\theta)/P(D)$

$$\propto P(D|\theta)*P(\theta)$$

Now we can incorporate our belief about θ

This is a MAP (Maximum A Posteriori) Estimate

Conjugate Prior

- Our likelihood so far has been based on a Bernoulli distribution.
- Beta is a conjugate prior to Bernoulli
  - This means their pdfs play nice together
  - P(D|θ)*P(θ) will be easy to deal with
  - Called the posterior likelihood

Beta/Binomial Distributions

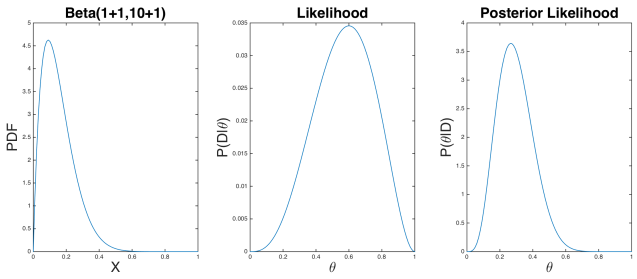
- Beta

$$P(\theta) \propto (1-\theta)^{\beta_p - 1} * \theta^{\beta_r - 1}$$

proportional to (missing a constant)

- Binomial

$$P(\theta) = (1-\theta)^{\alpha_p} * \theta^{\alpha_r}$$



More math

$$P(\theta|D)P(\theta) \propto [(1-\theta)^{\alpha_p} * \theta^{\alpha_r}] * [(1-\theta)^{\beta_p - 1} * \theta^{\beta_r - 1}]$$

Exercises:

1. find the optimal θ
2. What if $\beta_p = 1 \quad \text{and} \quad \beta_r = 1$?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given **prior probability and the data**

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

A wonderful tutorial:
http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf