Probability

Why Probability?

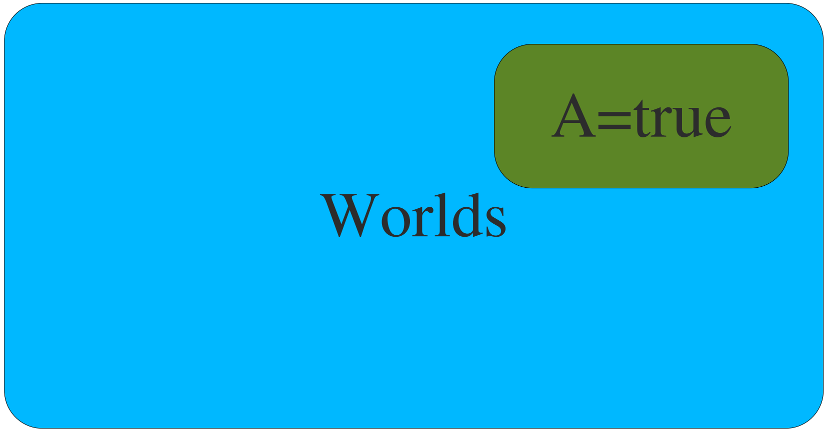
- It's the right way to look at the world.

Discrete Random Variables

- We denote discrete random variables with capital letters.
- A boolean random variable may be either true or false
  - A = true or A=false.
- P(a), or P(A=true) denotes the probability that A is true.
- Also unconditional probability or prior probability.
- P(¬a) or P(A=false) denotes the probability that A is not true.

Discrete Random Variables

- P(a): the fraction of worlds in which A is true.



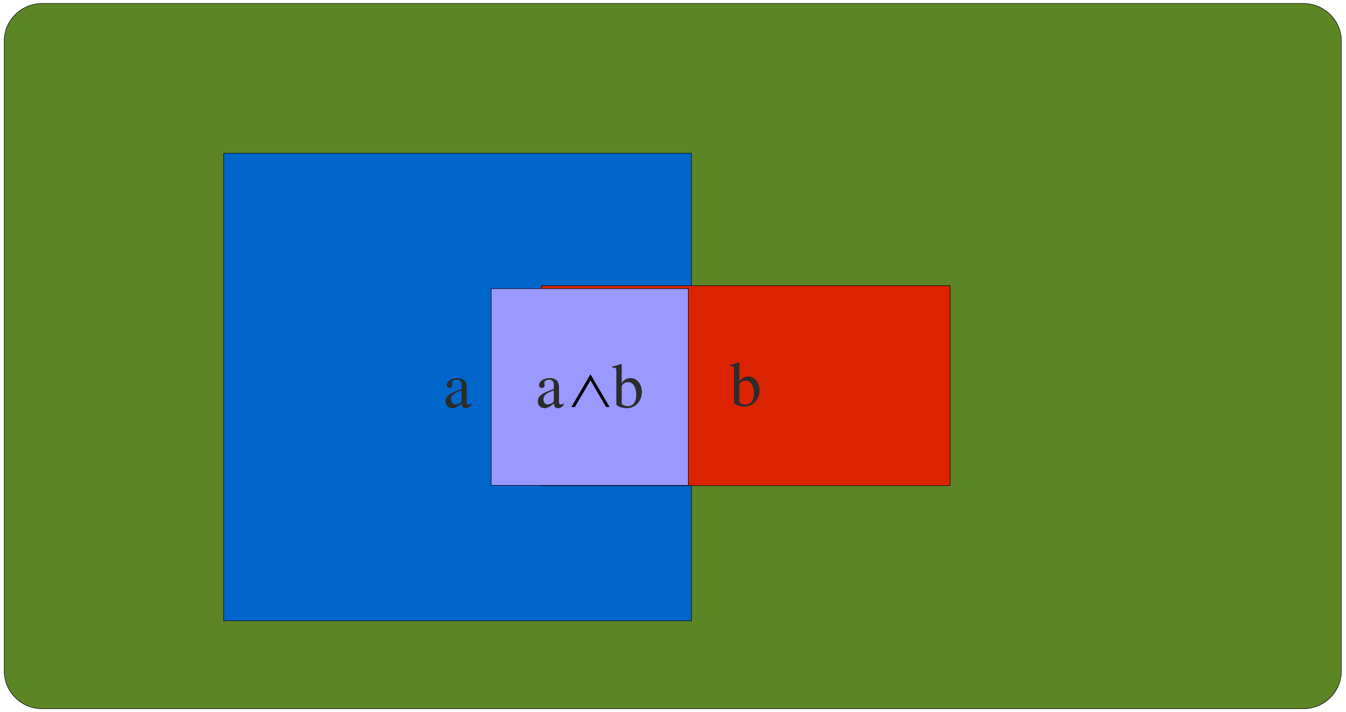
- $P(a) = .2 \quad P(\neg a) = .8$

More Notation

- We can apply boolean operators
  - Probability of a AND b: $P(a \wedge b)$ or $P(a,b)$
  - Probability of a OR b: $P(a \vee b)$

The Axioms of Probability

- $0 \leq P(a) \leq 1$
- $P(true) = 1 \quad P(false) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



A Simple Proof

- $0 \leq P(a) \leq 1$
- $P(true) = 1$, $P(false) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

- Prove that $P(\neg a) = 1 - P(a)$
  - $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$
  - $P(true) = P(a) + P(\neg a) - P(false)$
  - $1 = P(a) + P(\neg a) - 0$
  - $P(\neg a) = 1 - P(a)$

Multi-valued Random Variables

- We can define a random variable that can take on more than two possible values.
- E.g. C is one of $\{v_1, v_2, \ldots, v_N\}$
- Note, it must be the case that :

$$\sum_{1}^{N} P(v_i) = 1$$

- Example: W may have the domain {sunny, cloudy, rainy}.

Conditional Probability

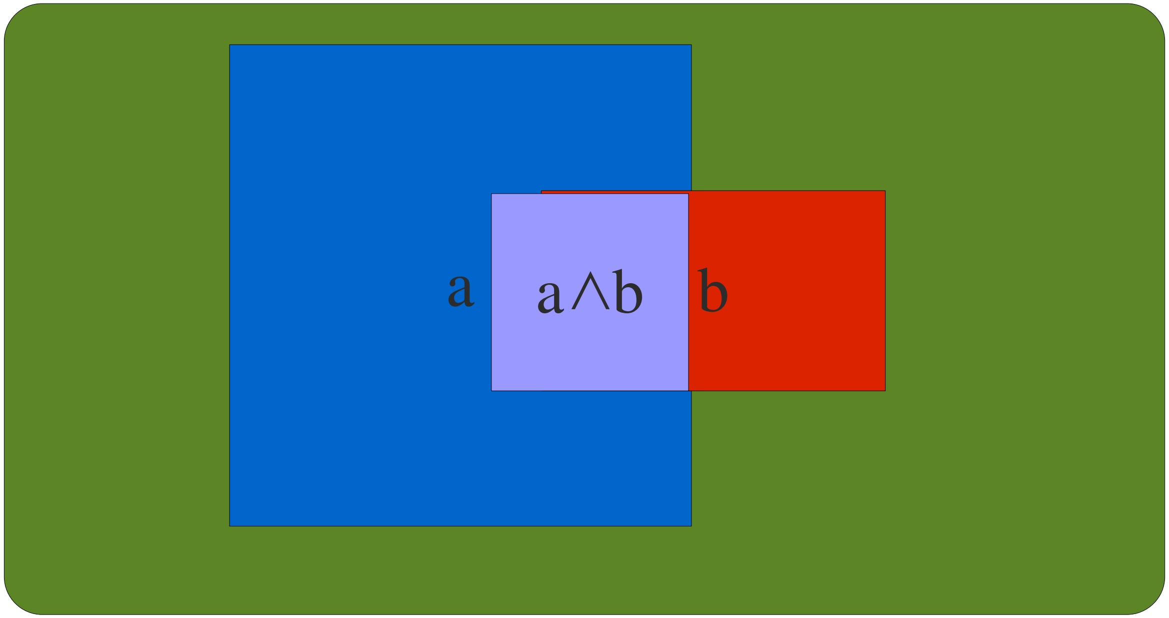
- P(a | b), the probability that A is true given that B is true.
  - P(sunny) = .1
  - P(sunny | warm) = .3
- Definition:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

- The fraction of worlds in which B is true, that also have A true.
- May also be written as the product rule:

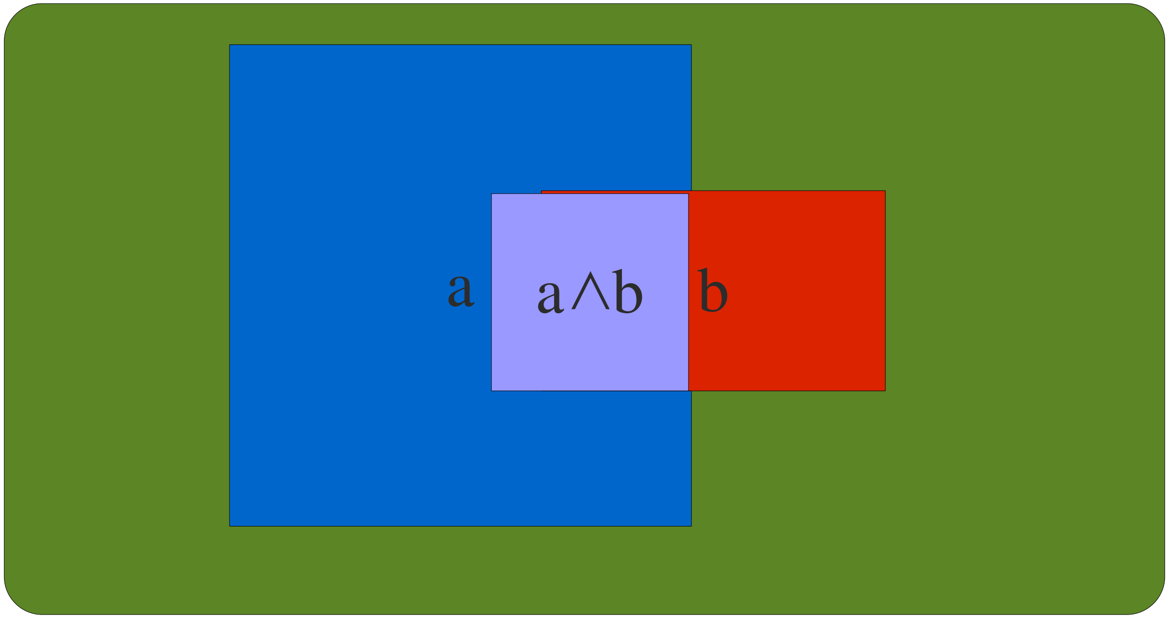
$$P(a \wedge b) = P(a|b)\,P(b)$$

Conditional Probability



- $P(a) = .3$
- $P(b) = .1$
- $P(a \wedge b) = .05$
- $P(a \mid b) = ??$

Conditional Probability



- $P(a) = .3$
- $P(b) = .1$
- $P(a \wedge b) = .05$
- $P(a \mid b) = .5$

Probability Distributions

- A probability distribution is a complete description of the probability of all possible assignments to a random variable.
- Examples:
- For a boolean variable
  - P(A=TRUE) = .1
  - P(A=FALSE) = .9
- Random variable W from the domain {sunny, cloudy, rainy}
  - **P**(W) = <.2 , .7, .1>

Joint Probability Distribution

- A complete description of the probability of all possible assignments to all variables (atomic event).

Two boolean variables A and B

A	B	Prob
T	T	.1
T	F	.2
F	T	.5
F	F	.2

Rooster Crows (C) and Weather (W)

C	W	Prob
T	sunny	.05
T	cloudy	.2
T	rainy	0
F	sunny	.05
F	cloudy	.4
F	rainy	.3

Inference

- Determining the probability of an event of interest, given everything that we know about the world.
- This is easy if we have the joint probability distribution.
- *The probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds.*

$$P(a) = \sum_{e_i \in \boldsymbol{e}(a)} P(e_i)$$

Inference Example

A	B	Prob
T	T	.1
T	F	.2
F	T	.5
F	F	.2

- What is P(A = true)?
- P(a) = ??

Inference Example

A	B	Prob
T	T	.1
T	F	.2
F	T	.5
F	F	.2

- What is P(A = true)?
- P(a) = .1 + .2 = .3
- In general $\boldsymbol{P}(Y) = \sum_{z} \boldsymbol{P}(Y, z)$ marginalization.

- Here Y and Z may be sets of variables, and the sum is over all possible assignments to the variables Z.

Conditional Inference

C	W	Prob
T	sunny	.05
T	cloudy	.2
T	rainy	0
F	sunny	.05
F	cloudy	.4
F	rainy	.3

- P(C=true | W =sunny)?
- Remember that: $$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$
- P(C=true | W = sunny) = ??

Conditional Inference

C	W	Prob
T	sunny	.05
T	cloudy	.2
T	rainy	0
F	sunny	.05
F	cloudy	.4
F	rainy	.3

- P(C=true | W =sunny)?
- Remember that: $$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$
- P(C=true | W = sunny) = .05 / (.05 + .05) = .5

"Learning" a Joint Probability Distribution

- Where does the joint probability distribution come from?
- Maybe we (or an expert) make it up.
- Or we can learn it: $\hat{P}(row) = \dfrac{\#\ instances\ that\ match\ row}{\#\ total\ instances}$

C	W	#days	Prob
T	sunny	12	12/38 =.32
T	cloudy	3	3/38 =.08
T	rainy	0	0/38 =.0
F	sunny	8	8/38 =.21
F	cloudy	10	10/38 =.26
F	rainy	5	5/38 =.13

total: 38

ANY PROBLEMS?

Problems with Learning PD

- This will quickly break down if we have more than a few variables.

Independence

- Variables A and B are independent if P(a | b) = P(a)
- We can also write: $P(a \wedge b) = P(a)P(b)$
  - Remember the product rule: $P(a \wedge b) = P(a|b)P(b)$
- Independence is a big deal for probabilistic reasoning.
  - Specifying the full joint PD requires exponential storage.
  - Learning it requires an exponentially growing amount of data.
  - These both become linear if all variables are independent.
  - This is called factoring the joint distribution.

Bayes' Rule

- The most useful identity in AI:

$$P(h|d)=\frac{P(d|h)P(h)}{P(d)}$$

- Think of h as hypothesis and d as data.
- $P(d|h)$ is called likelihood.

$$posterior = \frac{likelihood \times prior}{evidence}$$

- Why would we know P(d | h) and not P(h | d)?

Diagnosis

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

- I have a cough, I want to know the probability that I have pneumonia.
- P(cough) = .1, P(pneumonia) = .001, P(cough | pneumonia) = .5
- P(pneumonia | cough) = (.5 * .001) / .1 = .005 = .5%

(Simplistic) Spam Filtering

SPAM

viagra	discount	cs444	count
T	T	T	0
T	T	F	180
T	F	T	0
T	F	F	1200
F	T	T	8
F	T	F	600
F	F	T	12
F	F	F	6000
Total:			8000

NON-SPAM

viagra	discount	cs444	count
T	T	T	0
T	T	F	0
T	F	T	1
T	F	F	3
F	T	T	6
F	T	F	20
F	F	T	70
F	F	F	700
Total:			800

Bayes' Classifier I

- Assume a multivalued random variable C that can take on the values c_i for $i = 1$ to $i=K$.
- Assume M input attributes X_j for $j = 1$ to M.
- Learn $P(X_1, X_2, \ldots, X_M \mid c\_i)$ for each i.
  - Treat this as K different joint PDs.
- Given a set of input values $(X_1=u_1, X_2=u_2, \ldots, X_M=u_M)$, classification is easy (?):

$$C^{predict} = \underset{c_i}{argmax}\, P(C=c_i | X_1=u_1, X_2=u_2, \ldots, X_M=u_m)$$

An Aside: MAP vs. ML

- This is a maximum a posteriori (MAP) classifier:

$$C^{predict} = \underset{c_i}{argmax}\, P(C=c_i | X_1=u_1, X_2=u_2, \ldots, X_M=u_m)$$

- We could also consider a maximum likelihood (ML) classifier:

$$C^{predict} = \underset{c_i}{argmax}\, P(X_1=u_1, X_2=u_2, \ldots, X_M=u_m | C=c_i)$$

An Aside: Conditioning

- Remember marginalization?

$$\boldsymbol{P}(Y) = \sum_z \boldsymbol{P}(Y, z)$$

- We also have conditioning:

$$\boldsymbol{P}(Y) = \sum_z \boldsymbol{P}(Y|z)\boldsymbol{P}(z)$$

Bayes' Classifier II

$$C^{predict} = \underset{c_i}{argmax}\, P(C = c_i | X_1 = u_1, X_2 = u_2, \ldots, X_M = u_m)$$

- Apply Bayes' rule

$$C^{predict} = \underset{c_i}{argmax} \frac{P(X_1 = u_1, X_2 = u_2, \ldots, X_M = u_m | C = c_i) P(C = c_i)}{P(X_1 = u_1, X_2 = u_2, \ldots, X_M)}$$

- Conditioning:

$$C^{predict} = \underset{c_i}{argmax} \frac{P(X_1 = u_1, X_2 = u_2, \ldots, X_M = u_m | C = c_i) P(C = c_i)}{\sum_{i=1}^{K} P(X_1 = u_1, X_2 = u_2, \ldots, X_M | C = c_i) P(C = c_i)}$$

Bayes' Classifier III

$$C^{predict} = \underset{c_i}{argmax} \frac{P(X_1 = u_1, X_2 = u_2, \ldots, X_M = u_m | C = c_i) P(C = c_i)}{\sum_{i=1}^{K} P(X_1 = u_1, X_2 = u_2, \ldots, X_M | C = c_i) P(C = c_i)}$$

- Notice that the denominator is the same for all classes.
- We can simplify this to:

$$C^{predict} = \underset{c_i}{argmax}\, P(X_1 = u_1, X_2 = u_2, \ldots, X_M = u_m | C = c_i) P(C = c_i)$$

- If your learned distributions are correct, this is the <u>best</u> choice.
- What's the problem?

Naïve Bayes' Classifier

- If M is largish it is impossible to learn $P(X_1, X_2, \ldots, X_M \mid c\_i)$.
- The solution (?): assume that the X_j are independent given C (that the symptoms are independent, given the disease.)

$$P(X_{1,} X_{2,} \ldots, X_M | c_i) = \prod_{j=1}^{M} P(X_j | c_i)$$

- Factorization!
- The naïve Bayes' classifier:

$$C^{predict} = \underset{c_i}{argmax}\, P(C = c_i) \prod_{j=1}^{M} P(X_j | c_i)$$

Why is that Naïve?

- The symptoms probably *aren't* independent given the disease.
- Assuming they are allows us to classify based on thousands of attributes.
- This seems to work pretty well in practice.

An Note on Implementation

- if M is largish this product can get really small. Too small.

$$C^{predict} = \underset{c_i}{argmax}\, P(C=c_i) \prod_{j=1}^{M} P(X_j|c_i)$$

- Solution:

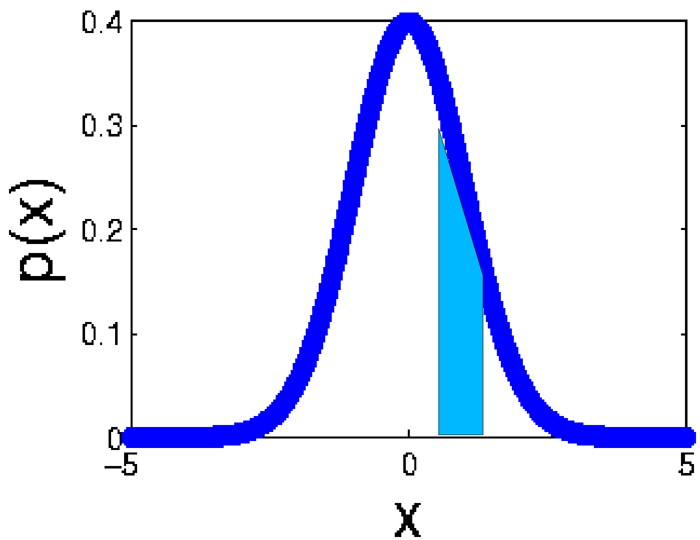
$$C^{predict} = \underset{c_i}{argmax} \left(\log P(C=c_i) + \sum_{j=1}^{M} \log P(X_j|c_i) \right)$$

- Remember that log(ab) = log(a) + log(b)

Continuous Random Variables

- Let X be a continuous random variable.
- if p(x) is a probability density function for X then:

$$P(a < X \le b) = \int_a^b p(x)\,dx$$



- The probability of any particular x is 0.

Probability Density Functions

- An equivalent definition:

$$p(x) = \lim_{h \to 0} \left(\frac{P(x - \frac{h}{2} < X \leq x + \frac{h}{2})}{h} \right)$$

- The ratio of the probability of landing in a region h, over the area of the region h approaches p(x) as h approaches 0.

Joint Probability Density Functions

- Consider two random variables X and Y and a two dimensional region R:

$$P((X,Y)\in R)=\int\int_{(x,y)\in R} p(x,y)\,dy\,dx$$

- The volume of the region R bounded above by p(x,y) corresponds the the probability that X and Y will be in R.