Overview of Probability Theory

Adaptead from

http://www.cs.cmu.edu/~tom/10701_sp11/slides/Overfitting_ProbReview-1-13-2011-ann.pdf

Probability Overview

- **Events**
  - Discrete random variables,
  - continuous random variables,
  - compound events.
- **Axioms of probability**
  - What defines a reasonable theory of uncertainty
- **Independent events**
- **Conditional probabilities**
- **Bayes rule and beliefs**
- **Join probability distribution**
- **Expectations**
- **Independence, Conditional independence**

Random Variables

- **Informally , A is random variable if**
  - A denotes something about which we are uncertain
  - Perhaps the outcome of a randomized experiment

- **Examples**
  - A = True if randomly drawn person from our class is female
  - A = The hometown of a randomly drawn person from our class
  - A = True if two randomly drawn persons from our classes have same birthday

- **Define P(A)** as "*the fraction of possible worlds in which A s true*" **or** "the fraction of times A holds, in repeated runs of the random experiment"
  - The set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S

$$A : S \rightarrow \{0, 1\}$$

A Little Formalism

More formally, we have

- A sample space S (e.g., set of students in our class)
  - aka the set of possible worlds

- a **random variable** is a function defined over the sample space
  - Gender : S → {m,f}
  - Height : S → Reals

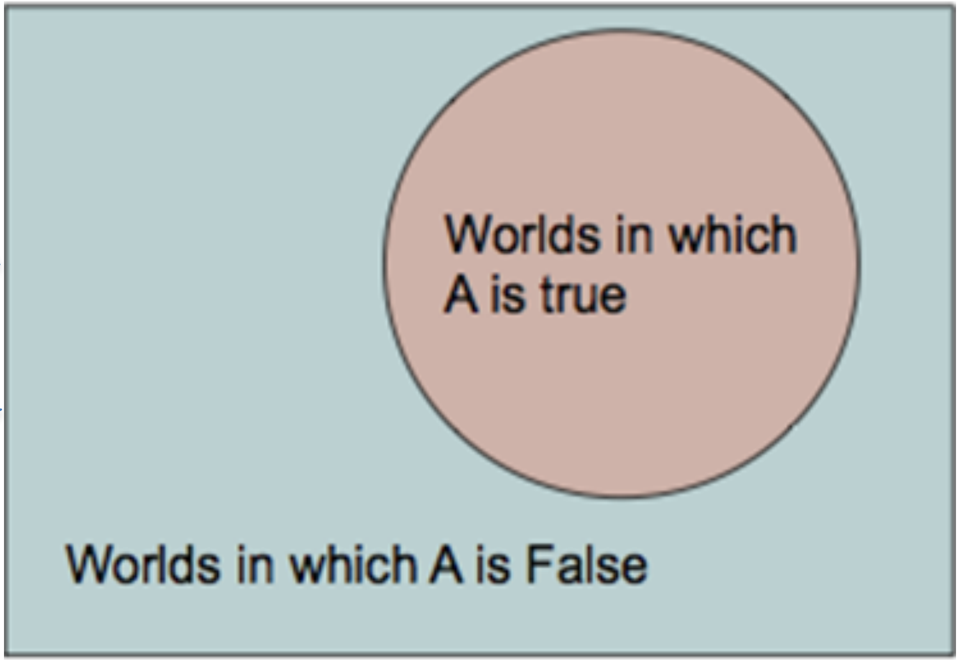
- an **event** is subset of S
  - e.g., the subset of **S** for which ***Gender=f***
  - e.g., the subset of **S** for which ***(Gender=m) AND (eyeColor=blue)***

- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

Visualizing A

Sample space of all possible worlds

Its area is 1



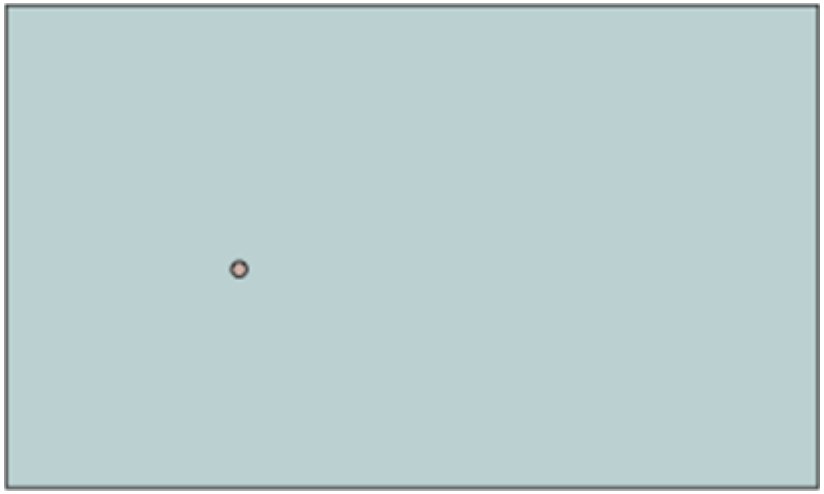
P(A) = Aa of reddish oval

The Axioms of Probability

- **$0 <= P(A) <= 1$**
- $P(True) = 1$
- $P(False)=0$
- **$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$**

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)**



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

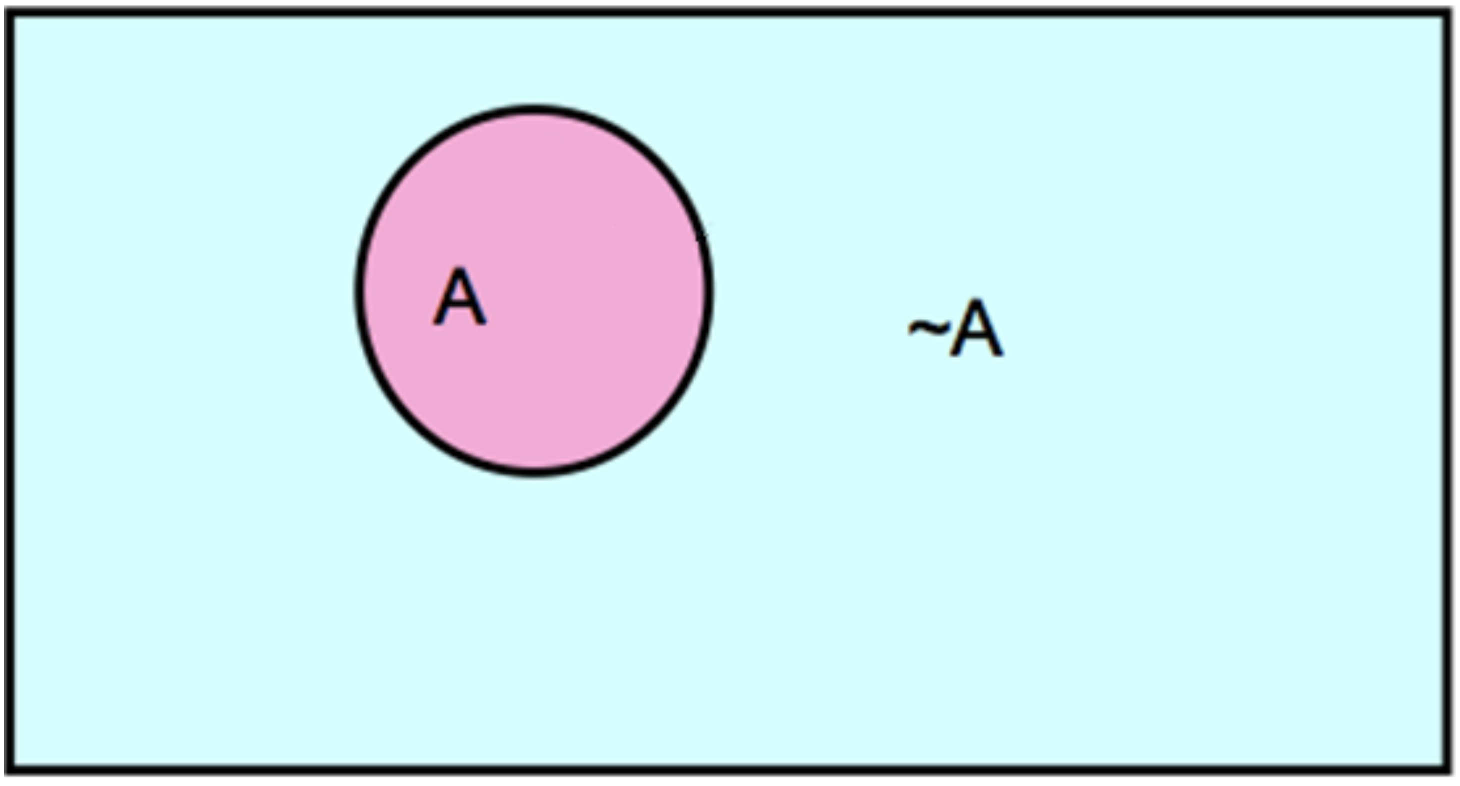
- **0 <=** P(A) <= 1
- P(True) = 1
- P(False)=0
- **P(A or B) = P(A) + P(B) - P(A and B)**

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

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

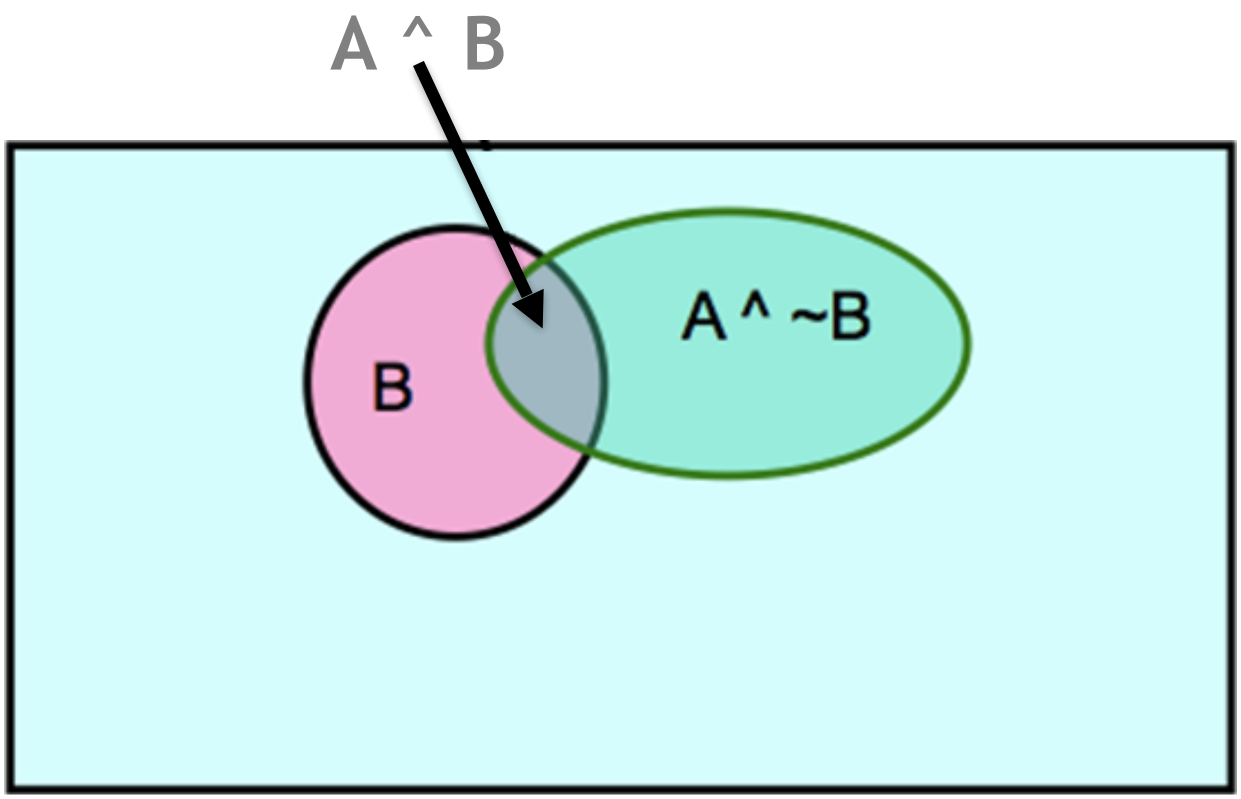
Elementary Probability in Pictures

- $P(\sim A) + P(A) = 1$



Elementary Probability in Pictures

- $P(A) = P(A \wedge B) + P(A \wedge \sim B)$



Multivalued Discrete Random Variables

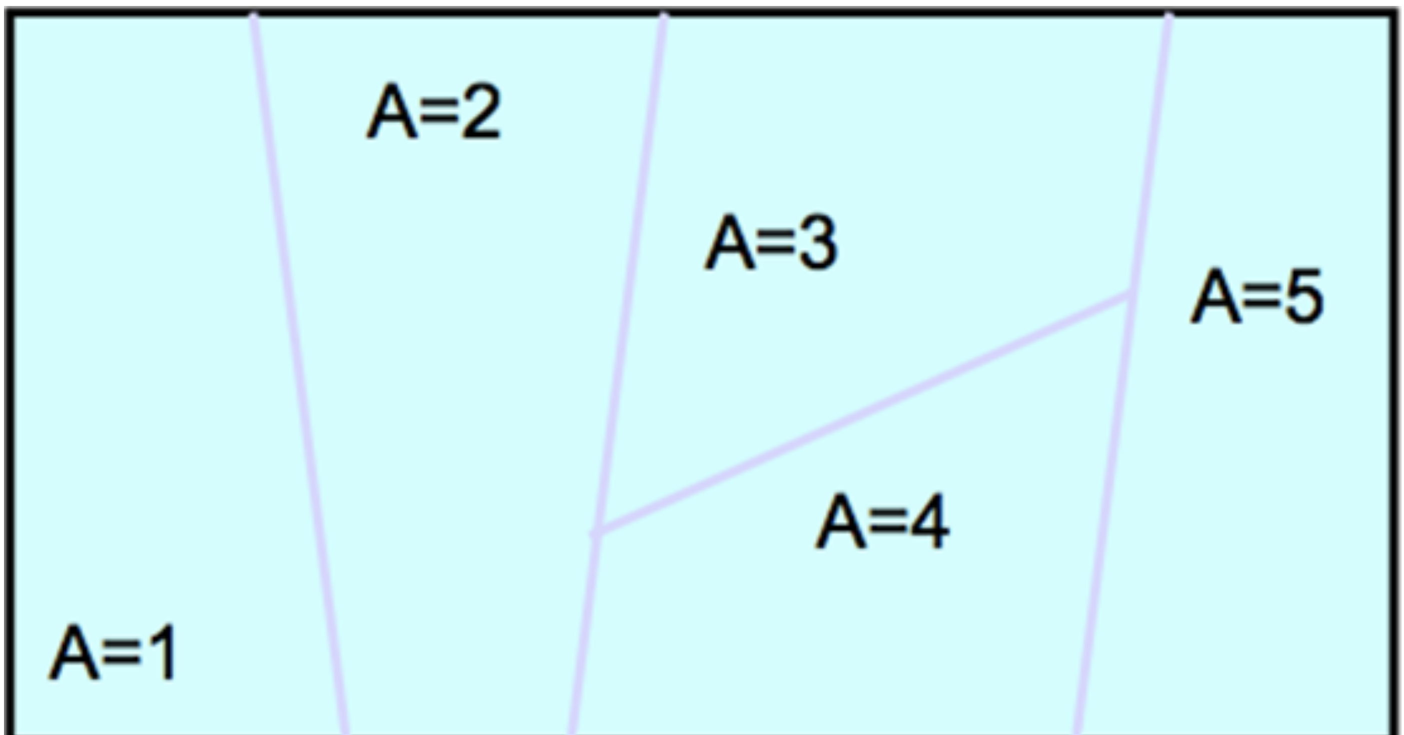
- Suppose A can take on more than 2 values
- A is random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus …

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

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

Elementary Probability in Pictures

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



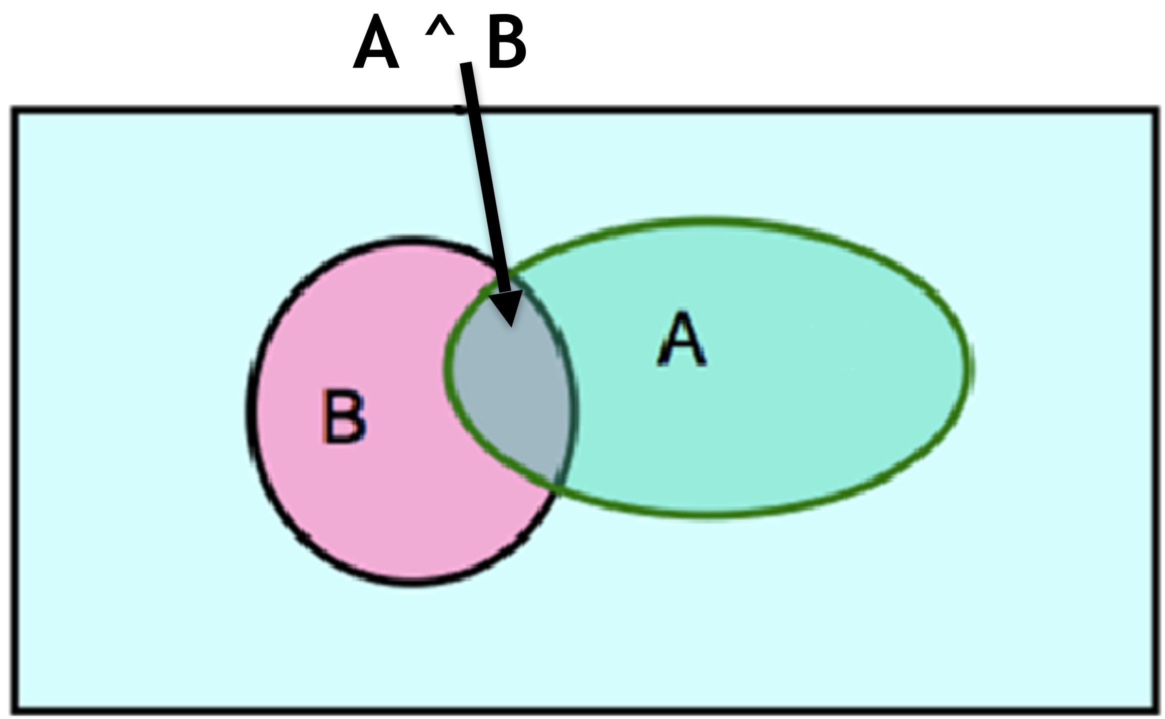
Independent Events

- Definition : two events **A** and **B** are *independent* if **Pr(A and B) = Pr(A) * Pr(B)**

- Intuition : knowing A tells us nothing about the value of B (and vice versa)

Visualizing Probabilities

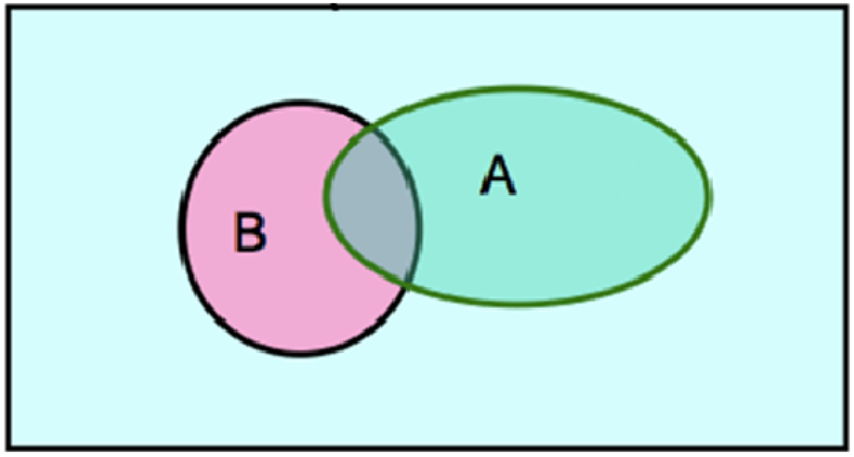
Sample space of all possible worlds



Its area is 1

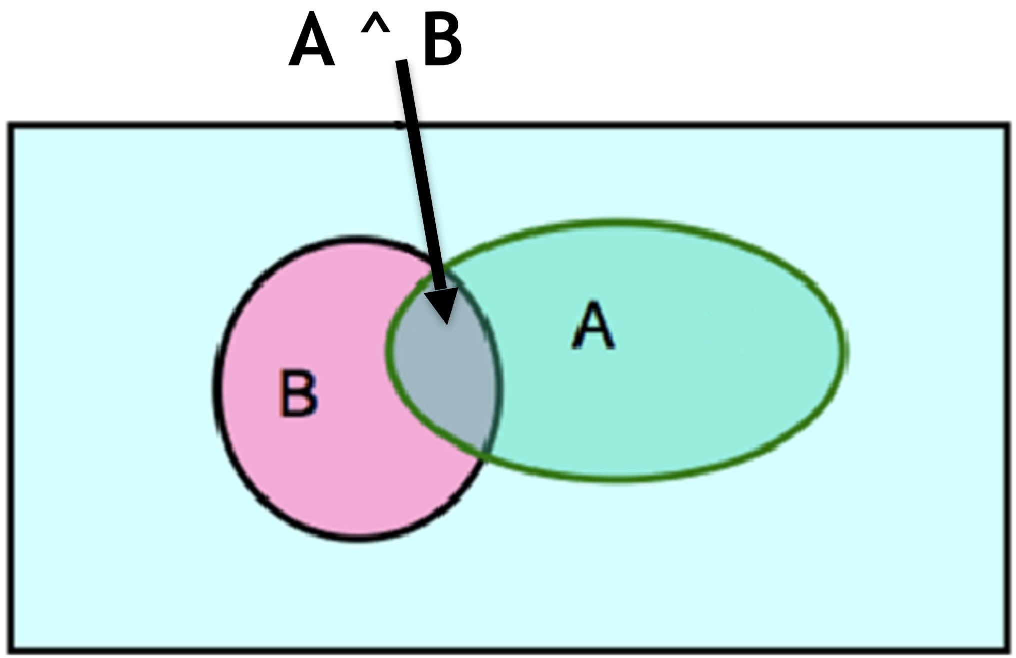
Definition of Conditional Probability

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



Bayes Rule

- Let's write 2 expressions for P(A ^ B)



Bayes' Rule

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

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 chance, but necessary to be solved in order to a sure foundation for all our reasoning 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 …

Definition of Conditional Probability

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

Corollary : The Chain Rule

P(A^B) = P(A|B) P(B)

P(C^A^B) = P(C|A^B) P(A|B) P(B)

Other Forms of Bayes Rule

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

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

Applying Bayes Rule

- A = You have the flu
- B = You just coughed
- The chance of getting flue is 0.05
- The chance of coughed given you have the flue is 0.8
- The chance of coughed given you do not have the flue is 0.2
- What is the chance of getting the flue given the fact the you just coughed

Applying Bayes Rule

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

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

Assume :
P(A) = 0.05
P(B|A)=0.80
P(B|~A)=0.2

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

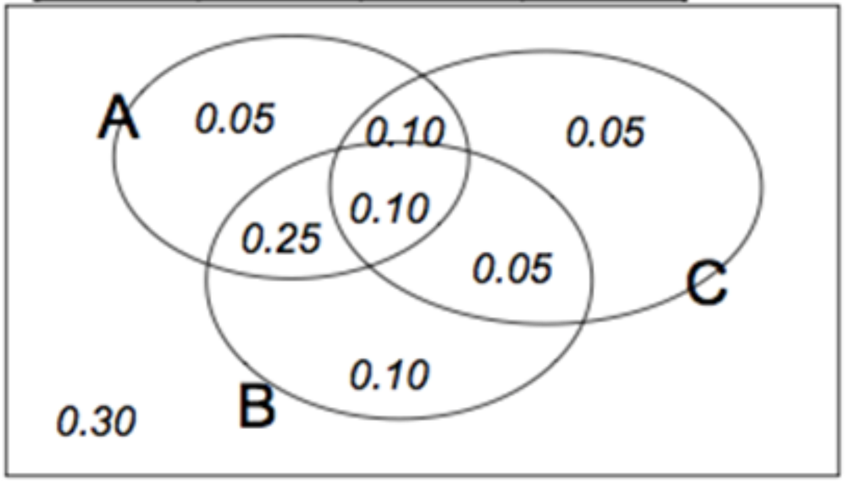
what does all this have to do with function approximation ?

The Joint Distribution

Recipe for making a joint distribution of M variables :

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



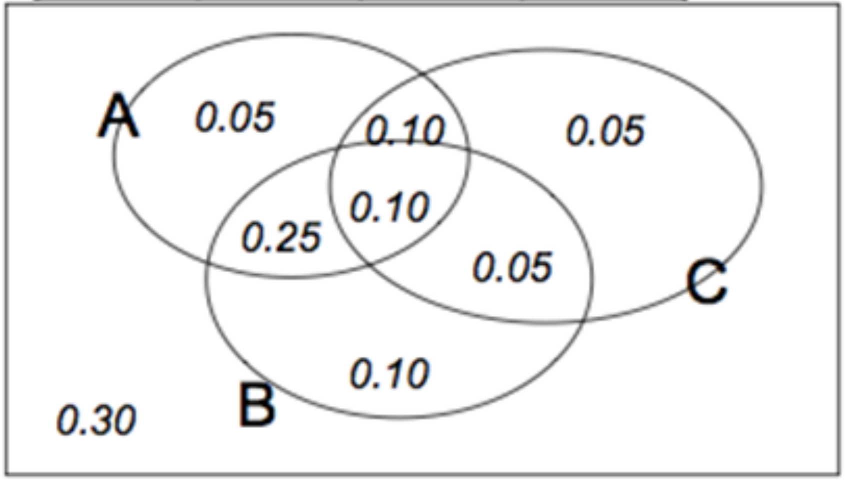
The Joint Distribution

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	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



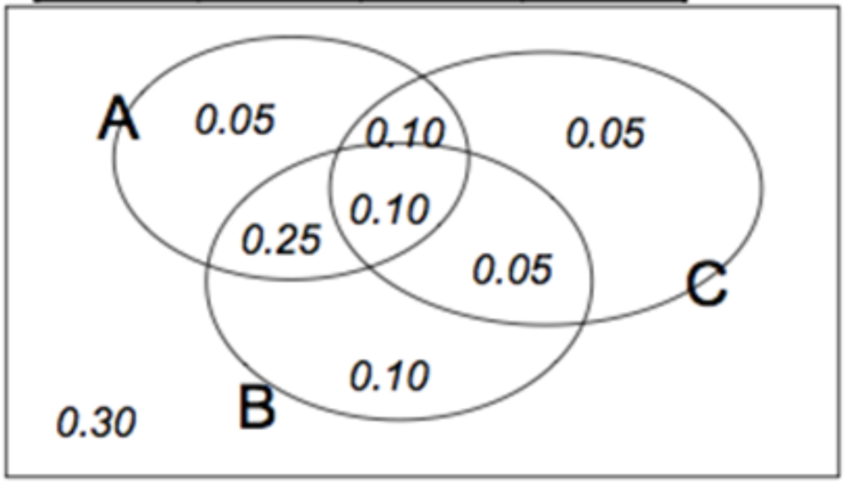
The Joint Distribution

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)
2. For each combination of values, say how probable it is.

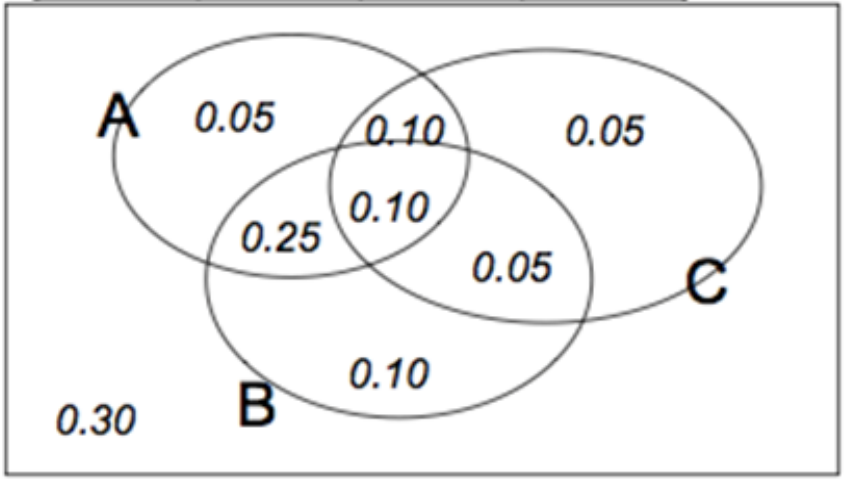
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



The Joint Distribution

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

$$P(E) = \sum_{rows \quad matching \quad E} P(ROW)$$

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

$$P(E) = \sum_{rows \; matching \; E} P(ROW)$$

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

$$P(E) = \sum_{rows \quad matching \quad E} P(ROW)$$

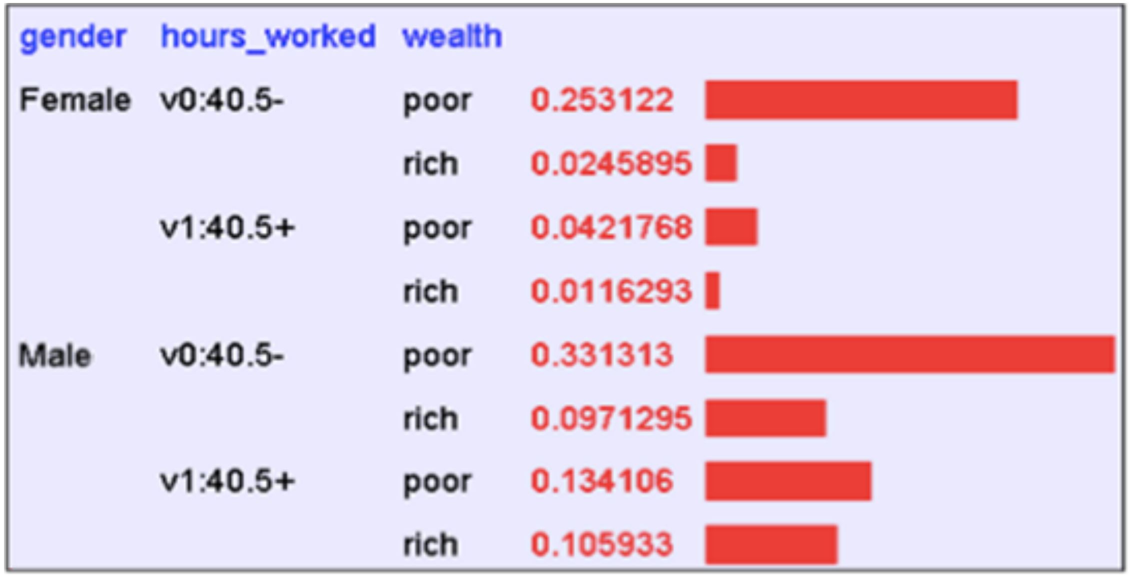
Inference with the Joint

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

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

P(Male | Poor) = ?

Learning and the Joint Distribution



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	

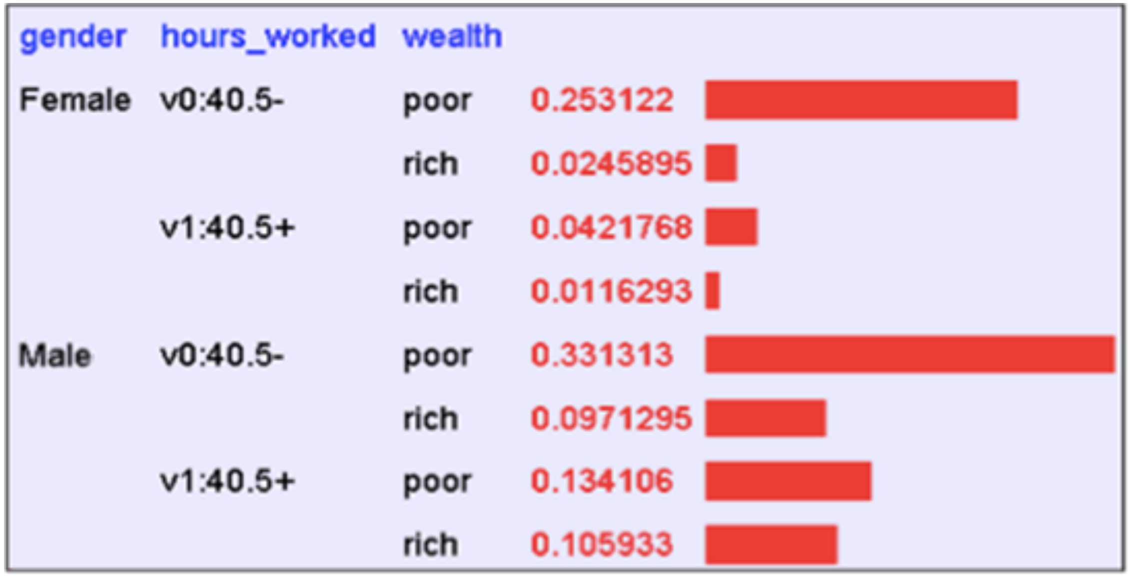
Suppose we want to learn the function f : <G, H> → W

Equivalently, P(W|G, H)

Solution : learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

Learning and the Joint Distribution



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	

e.g., P(W=rich | G = female, H = 40.5-) =

You should know

- Event
  - Discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs

sounds like the solution to learning $F: X \rightarrow Y$,

Or $P(Y|X)$.

Are we done ?

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question :
  - He says : I have thumbtack, if I flip it, what's the probability it will fall the nail up ?
  - You say : Please flip it a few times :

  - You say : The probability is :
  - **He says : Why ???**
  - You say : Because ….

Thumbtack - Binomial Distribution

- P(Heads) = θ , P() = 1 - θ

- Flips are i.i.d. :
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails.

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximun Likelihood Estimation

- Data : Observed set *D* of α_H Heads and α_T Tails.
- Hypothesis : Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function ?

- MLE : choose θ that maximize the probability of observed data :

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

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

Maximun Likelihood For θ

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

- Set derivation to zero : $\dfrac{d}{d\theta} \ln P(D \mid \theta) = 0$

Maximun Likelihood For θ

- **Set derivation to zero :** $$\frac{d}{d\theta}\ln P(D\,|\,\theta\,) = 0$$

$$\begin{aligned}
\widehat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \; \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}
\end{aligned}$$

Maximun Likelihood For θ

How many flips do I need ?

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Bayesian Learning

Use Bayes rule :

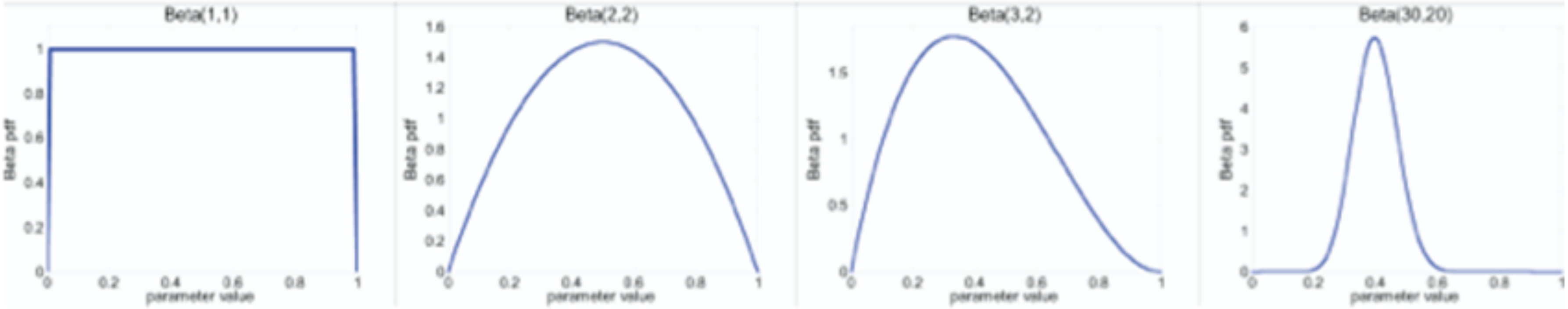
$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

Or equivalently

$$P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$$

Beta prior distribution - P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



- Likelihood function : $P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

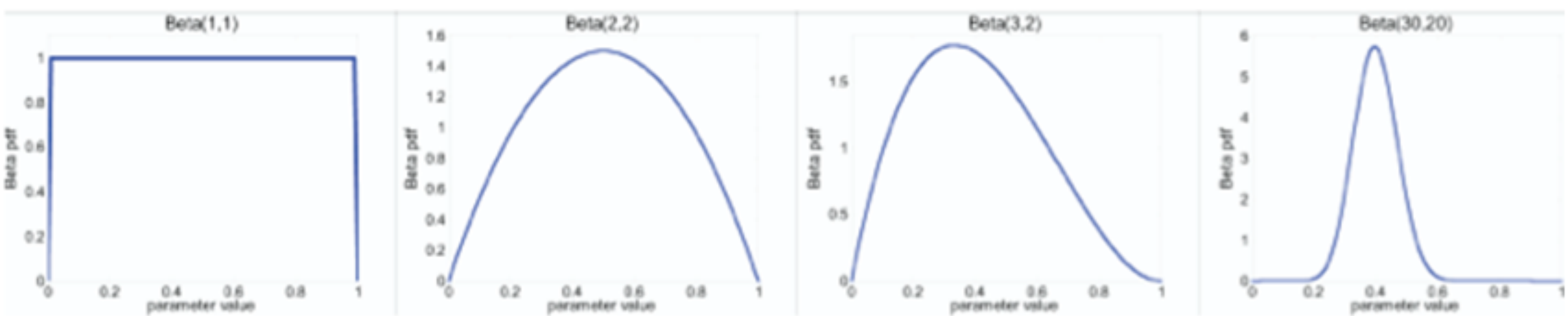
- Posterior : $P(\theta \mid D) \propto P(D \mid \theta) P(\theta)$

Posterior distribution

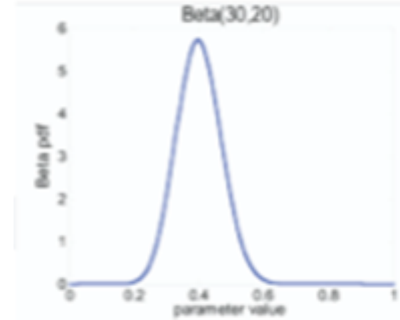
- **Prior :** $Beta(\beta_H, \beta_T)$
- **Data :** α_H **heads and** α_T **tails**

- **Posterior distribution:**

$$P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



MAP for Beta distribution



$$P(\theta \mid D) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- **MAP : use most likely parameter**

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

- **Beta prior equivalent to extra thumbtack flips**
- **As *N* → ∞, prior is "forgotten"**
- **But, for small sample size, prior is important!**

Estimating Parameters

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

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

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

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad P(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \quad = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}\end{aligned}$$

You should know

- **Probability basics**
  - random variables, events, sample space, conditional probs, …
  - Independence of random variables
  - Bayes rule
  - Joint probability distributions
  - Calculating probabilities from the joint distribution
- **Point estimation**
  - Maximum likelihood estimates
  - Maximum a posteriori estimates
  - Distributions – binomial, Beta, Dirichlet, …