Probability and Estimation

Donghai Guan
NUAA

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
- Joint probability distribution
- Expectations
- Independence, Conditional independence

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment
- Examples
  - A=True if a 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 class have same birthday
- Define P(A) as “the fraction of possible worlds in which A is 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

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

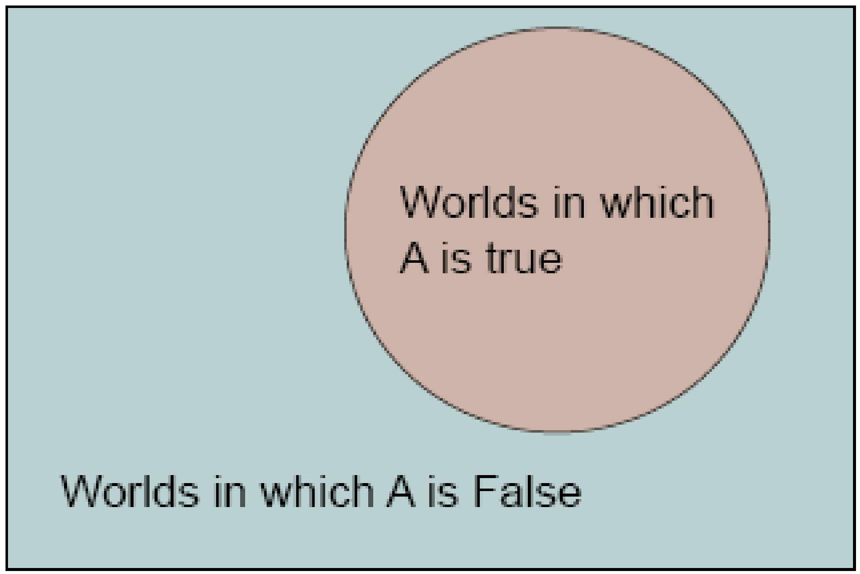
A little formalism

- 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 a 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 are often interested in probabilities of specific events
- And of specific events conditioned on other specific events

Visualizing A

Sample space of all possible worlds

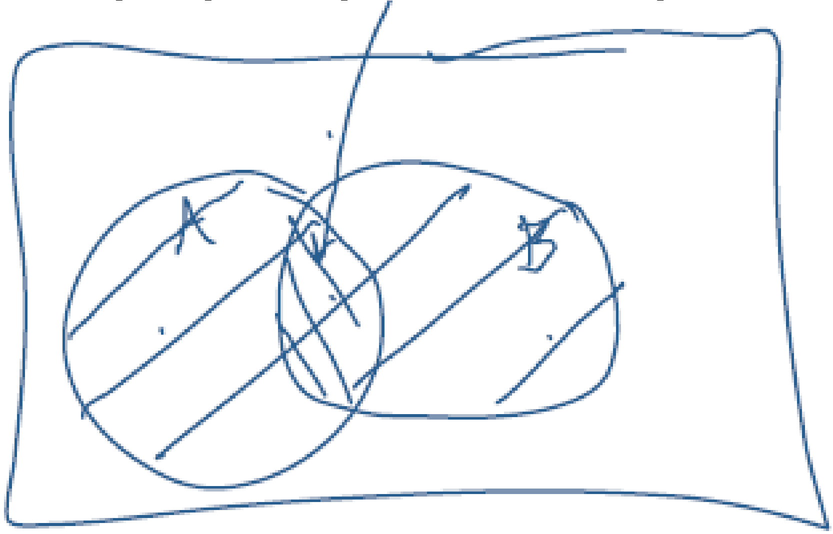
Its area is 1



P(A) = 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)



Interpreting the axioms

- The area of A can’t get any smaller than 0
- And a zero area would mean no world could ever have A true
- The area of A can’t get any larger than 1
- And an area of 1 would mean all worlds will have A true

Theorems from the Axioms

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

$\rightarrow$ $P(\text{not } A) = P(\sim A) = 1\text{-}P(A)$

$P(A \text{ or } \sim A) = 1$ $\qquad$ $P(A \text{ and } \sim A) = 0$

$$\underbrace{P(A \text{ or } \sim A)}_{1} = P(A) + P(\sim A) - \underbrace{P(A \text{ and } \sim A)}_{0}$$

$$1 = P(A) + P(\sim A) - 0$$

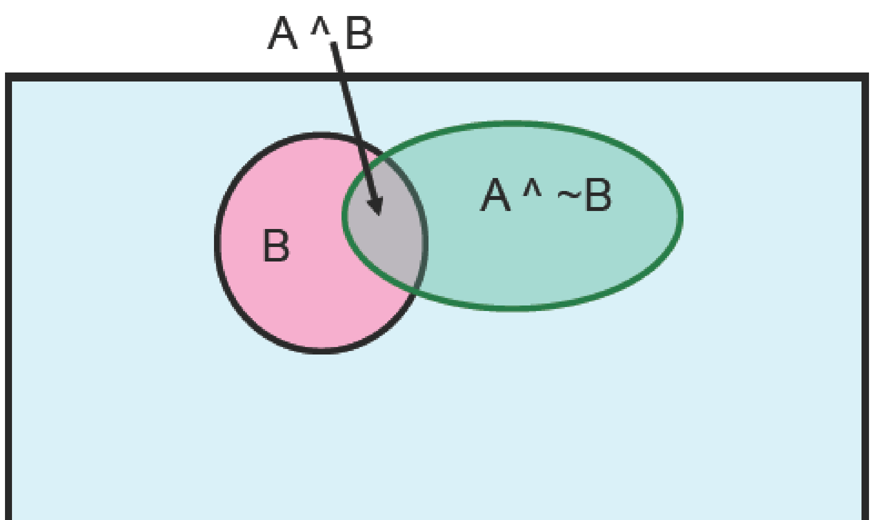
Elementary Probability in Pictures

- P(~A) + P(A) =1
- P(A)=P(A ^ B) + P(A ^ ~B)

A = A and (B or ~B) = (A and B) or (A and ~B)

P(A) = P(A and B) + P(A and ~B) – P((A and B) and (A and ~B))

P(A) = P(A and B) + P(A and ~B) – P(~~A and A and B and ~B)~~



Multivalue Discrete Random Variables

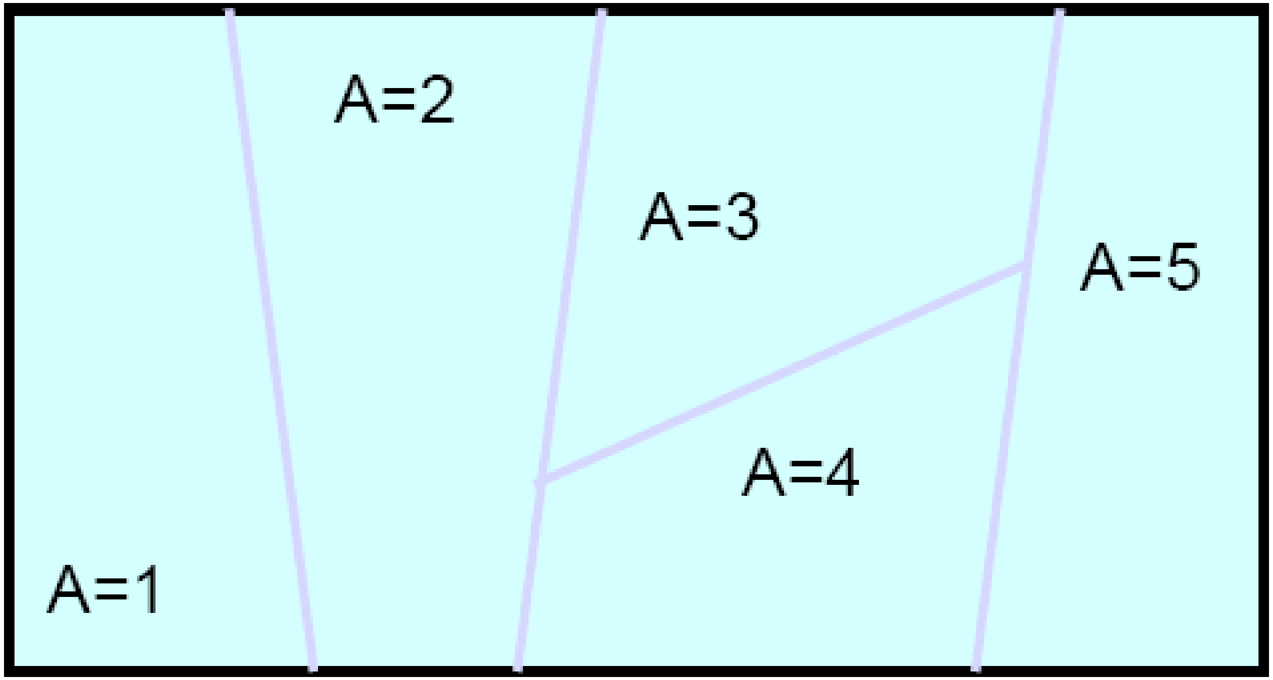
- Suppose A can take more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus…

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

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

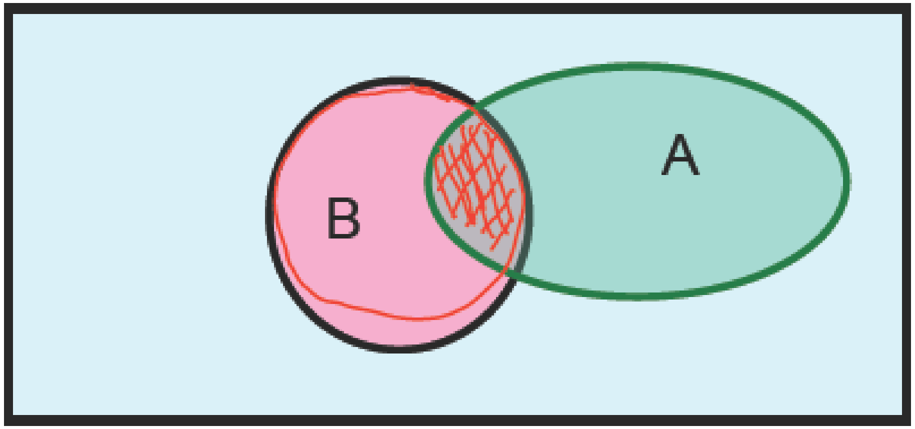
Elementary Probability in Pictures

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



Definition of Conditional Probability

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



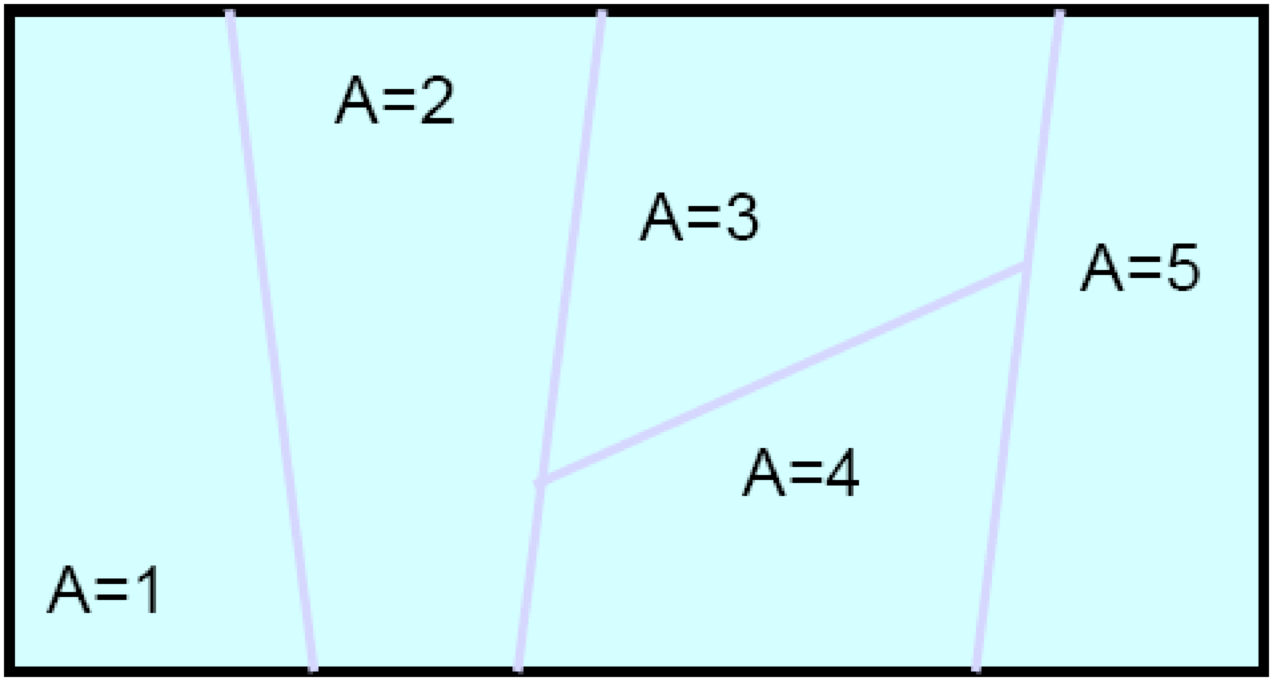
Corollary: The Chain Rule

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

$$P(C \wedge A \wedge B) = P(C \mid A \wedge B) P(A \mid B) P(B)$$

Conditional Probability in Pictures

picture: $P(B|A=2)$



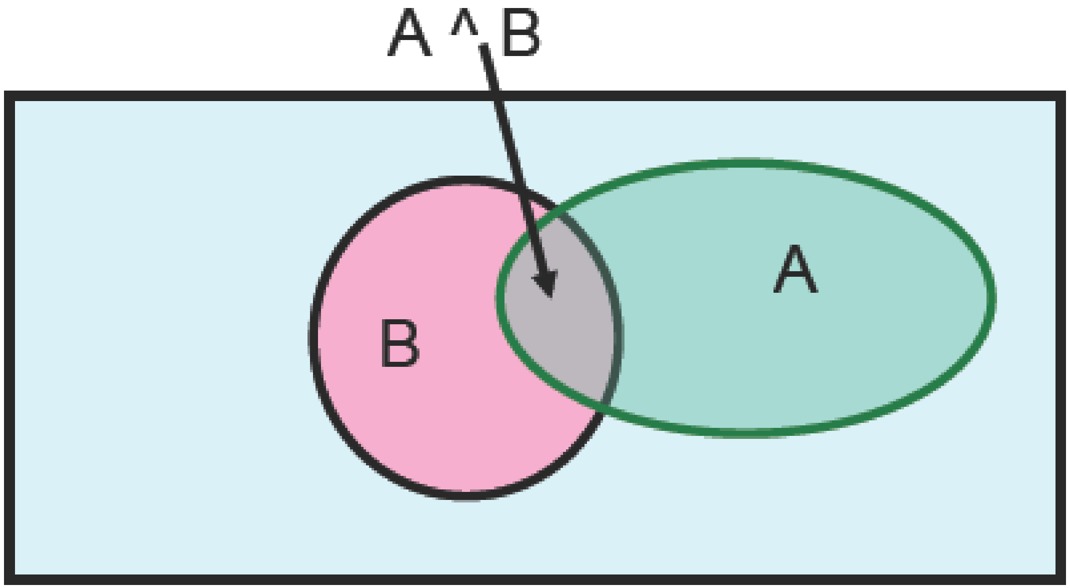
Independent Events

- 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)

Elementary Probability in Pictures

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



Bayes Rule

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

we call P(A) the "prior"

and P(A|B) the "posterior"

…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

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

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

Other Forms of Bayes Rule

$$P(A|BCD) = \frac{P(ABCD)}{P(BCD)} \quad (1)$$

$$P(A|BCD) = \frac{P(BCD|A)P(A)}{P(BCD)} \quad (2)$$

$$P(A|BCD) = \frac{P(B|ACD)P(A|CD)}{P(B|CD)} \quad (3)$$

$$P(A|BCD) = \frac{P(BC|AD)P(A|D)}{P(BC|D)} \quad (4)$$

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

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??

$f: X \rightarrow Y$

$P(Y|X)$

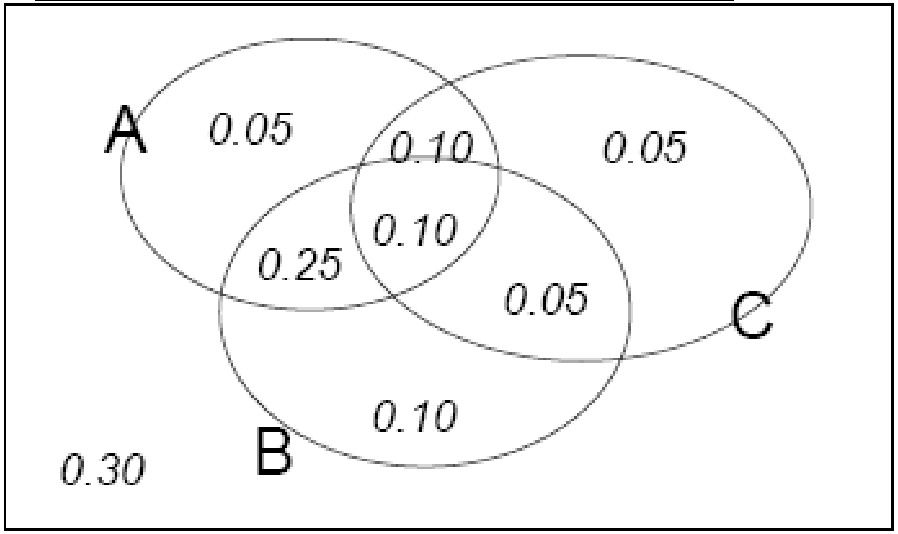
The Joint Distribution

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.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

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

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

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{rows\ \text{matching E}_1\text{ and E}_2} P(row)}{\sum_{rows\ \text{matching E}_2} P(row)}$$

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-) = $\frac{.024}{.024 + .25} < .1$

Sounds like the solution to learning

$F: X \rightarrow Y$, or $P(Y|X)$.

Are we done?

Your first consulting job

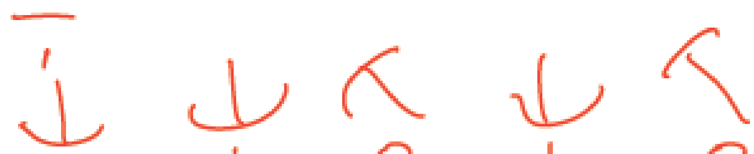
- A billionaire asks you a questions:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: please flip it a few times:

  - You say: The probability is: 0.6!
  - He says: Why???

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

α_T tails outcomes
α_H heads outcomes



D: $\{x_1, x_2, x_3, x_4, x_5\}$ $P(D|\theta)$

$P(D|\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) = \boxed{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}$

Flips produce data set D with α_H heads and α_T tails

-Flips are independent, identically distributed 1's and 0's (Bernoulli)

- α_H and α_T are counts that sum these outcome (Binomial)

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

Maximum Likelihood Estimation

- Data: Observed set D of α_H Heads and α_T Tails
- Hypothesis: Binomial distribution
- Learning θ is an optimization problem
  - What is the objective function?
- MLE: Choosing θ that maximizes the probability of observed data:

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

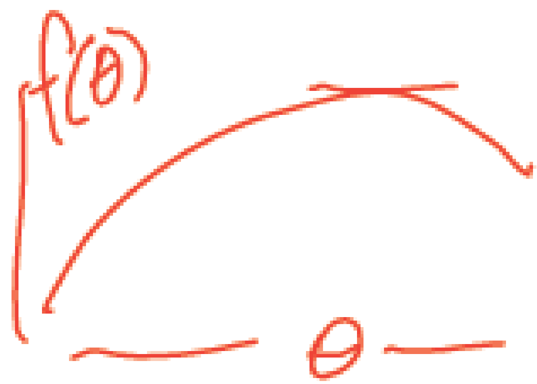
$$= \arg\max_{\theta} \ln P(D \mid \theta)$$

Maximum Likelihood Estimate for θ

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

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

$$= \arg\max_{\theta} \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$



- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

$$= \alpha_H \ln \theta + \alpha_T \ln(1-\theta)$$

$$\frac{d}{d\theta} \ln P(D \mid \theta) = \frac{d}{d\theta} (\alpha_H \ln \theta + \alpha_T \ln(1-\theta))$$

$$= \alpha_H \frac{1}{\theta} - \alpha_T \frac{1}{1-\theta} = 0$$

$$\Rightarrow \hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

MLE & MAP

Bayesian Learning

$MLE = \arg\max_{\theta} P(D|\theta)$

- Use Bayes rule:

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

Not dep. on θ

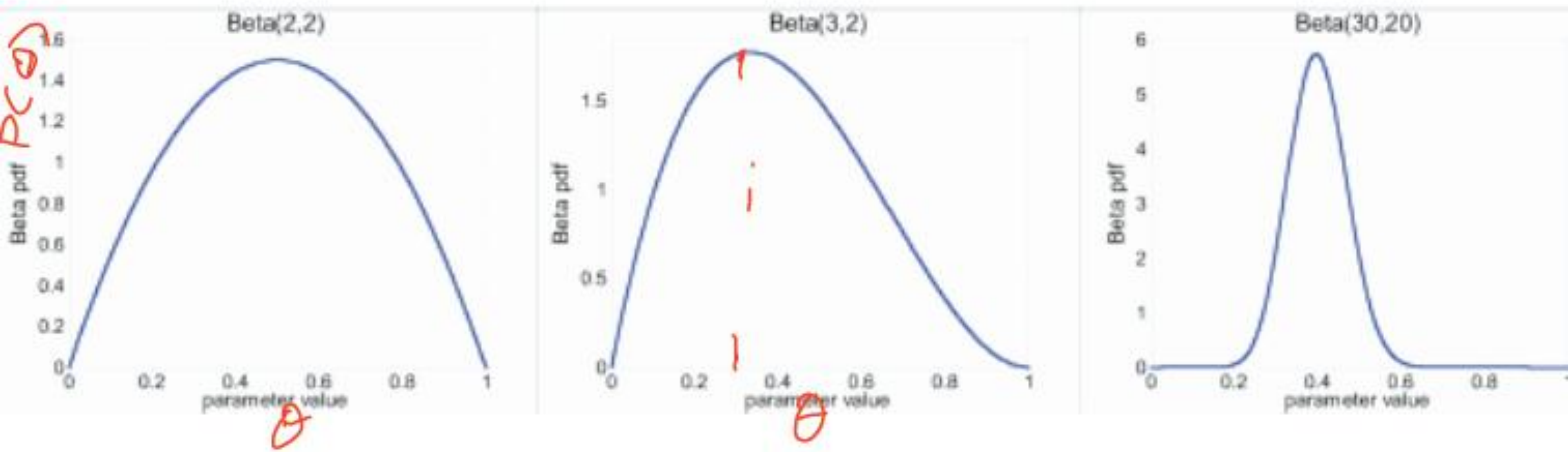
- Or equivalently:

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

$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$$

Beta prior distribution – $P(\theta)$

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



Beta prior distribution – $P(\theta)$

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

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

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

$$\theta^{\alpha_H}(1-\theta)^{\alpha_T} \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$$

$$= \frac{\theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}}{B()}$$

MAP

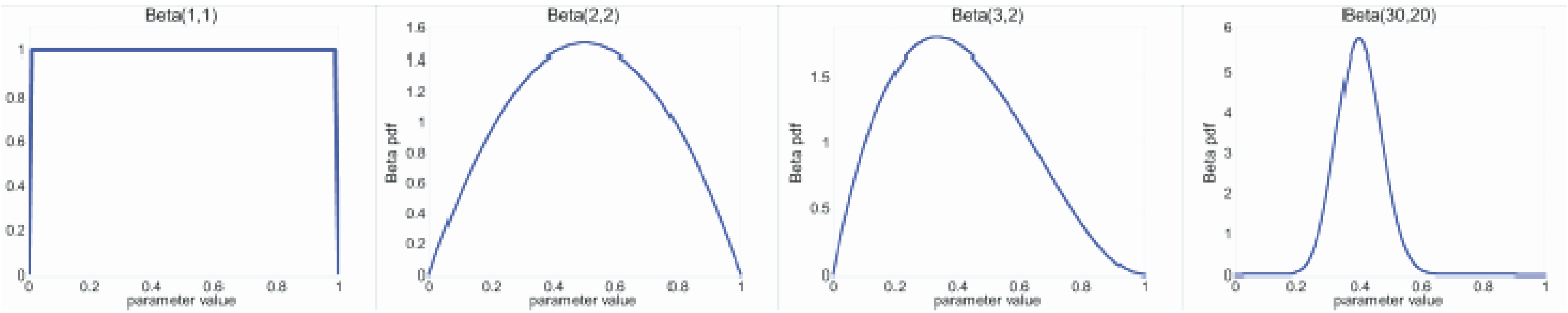
$$\arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} \frac{\theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}}{B()}$$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

Posterior distribution

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

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



MAP for Beta distribution

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

MAP: use most likely parameter: $\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$

- Beta prior equivalent to extra thumbtack flips
- As $N \to \infty$, prior is "forgotten"
- But for small sample size, prior is important!!

Conjugate priors

P(θ) and P(θ|D) have the same form

Eg.1 Coin flip problem

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

If Prior is Beta distribution, $P(\theta) = \dfrac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$

Then posterior is Beta distribution

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

For Binomial, conjugate prior is Beta distribution.

Conjugate priors

P(θ) and P(θ|D) have the same form

Eg.2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~Multinomial $P(D \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$

If Prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \beta_2, \ldots, \beta_k)} \sim Dirichlet(\beta_1, \beta_2, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta \mid D) = Dirichlet(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

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 Posterior (MAP) : choose θ that is most probable given prior probability and the data

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

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

Dirichelet distribution

- Number of heads in N flips of a two-sided coin
  - Follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)
- What it's not two-sided, but k-sided?
  - Follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \ldots, \theta_k) = \frac{1}{B(\alpha)} \prod_{i}^{K} \theta_i^{(\alpha_i - 1)}$$

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
- Estimating parameters from data
  - Maximum likelihood estimates
  - Maximum a posterior estimates
  - Distributions-binomial, Beta, Dirichlet, …
  - Conjugate priors

Expected values

- Given discrete random variable X, the expected value of X, written E[X] is

$$E[X] = \sum_{x \in X} xP(X = x)$$

- We also can talk about the expected value of functions of X

$$E[f(X)] = \sum_{x \in X} f(x)P(X = x)$$

Covariance

- Given two discrete r.v.'s X and Y, we define the covariance of X and Y as

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., X=gender, Y=playsFootball

Summary

- Questions?

- Thank You!