

CSCI 4520- Introduction to Machine Learning

Mehdi Allahyari
Georgia Southern University

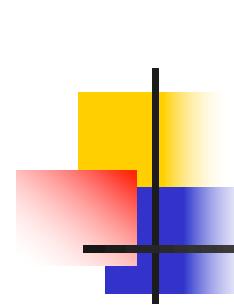
Basic Probability

(slides borrowed from Andrew Moore of CMU and Google,
<http://www.cs.cmu.edu/~awm/tutorials>, Tom Mitchell,
Barnabás Póczos & Aarti Singh



Probability Overview

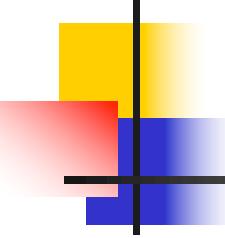
- Events
 - discrete random variables, continuous random variables, compound events
- Axioms of probability
 - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence
- MLE
- MAP



Discrete 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

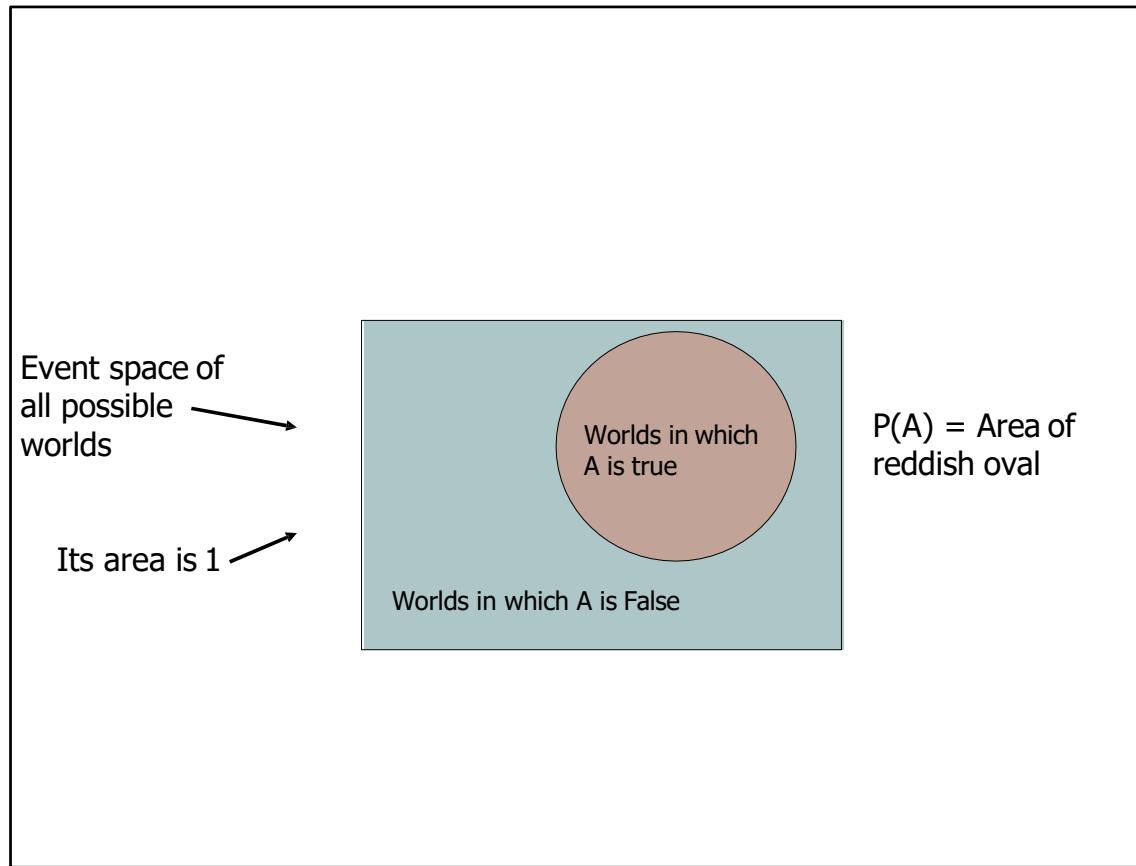
$$A: S \rightarrow \{0, 1\}$$

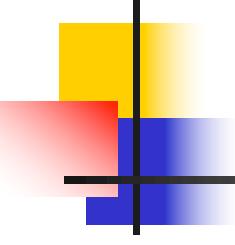


A little Formalism

- 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: $S \rightarrow \{ m, f \}$
 - Height: $S \rightarrow \text{Reals}$
 - 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're often interested in probabilities of specific events
 - and of specific events conditioned on other specific events

Visualizing A



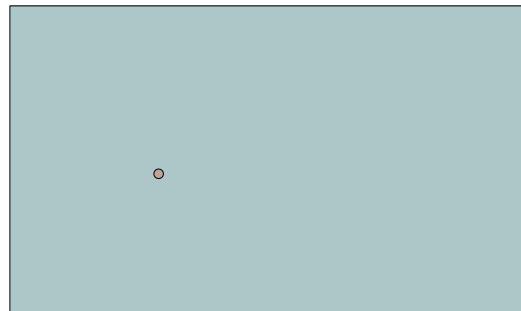


The Axioms of Probability

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

Interpreting the axioms

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

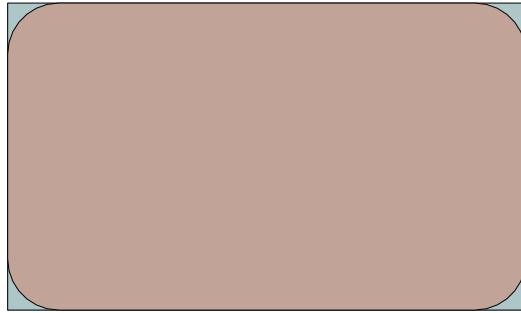


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

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

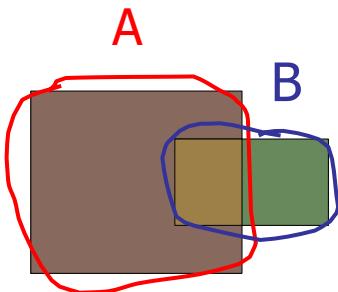


The area of A can't get any bigger than 1

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

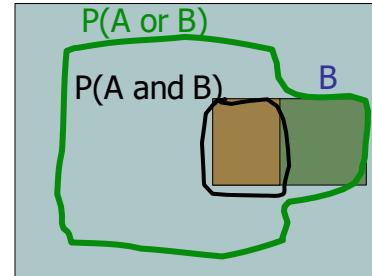
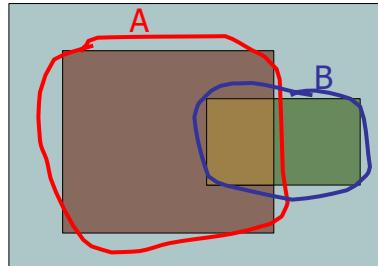
Interpreting the axioms

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

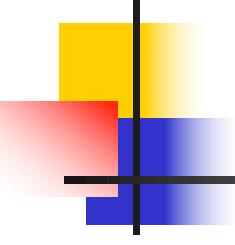


Interpreting the axioms

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



Simple addition and subtraction



Theorems from the Axioms

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

From these we can prove:

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

- How?

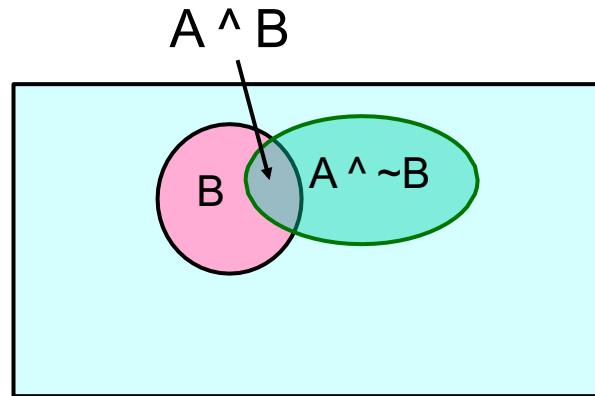
Another important theorem

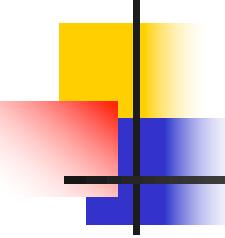
- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$

From these we can prove:

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

- How?





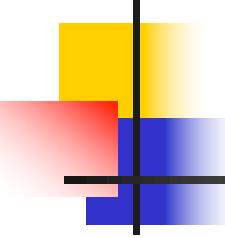
Notation Digression

Binary valued variables

- $P(A)$ is shorthand for $P(A=\text{true})$
- $P(\sim A)$ or $P(\bar{A})$ is shorthand for $P(A=\text{false})$
- For binary variables that aren't true/false:
 $P(\text{Gender}=\text{M})$, $P(\text{Gender}=\text{F})$

Multivalued variables

- $P(\text{Major}=\text{history})$, $P(\text{Age}=19)$, $P(Q<13.75)$



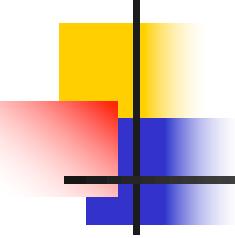
Notation Digression

Note: upper case letters/names for *variables*, lower case letters/names for *values*

- $P(X)$ vs $P(x)$
- For multivalued RVs, $P(Q)$ is shorthand for $P(Q=q)$ for some unknown q

Conjunctions

- $P(X, Y)$ equivalent to $P(X \wedge Y)$

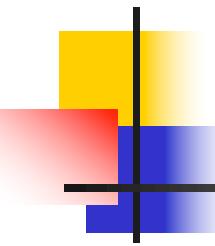


Multivalued Random Variables

- 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 $\{v_1, v_2, \dots, v_k\}$
- Thus...

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

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



An easy fact about Multivalued Random Variables:

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- And assuming that A obeys...

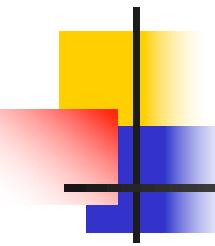
$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

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

- It's easy to prove that

- $P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = \sum_{j=1}^{j=k} P(A = v_j)$

Copyright © Andrew W. Moore



Another fact about Multivalued Random Variables:

- Using the axioms of probability...

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

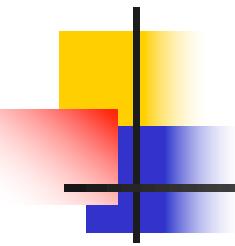
- And assuming that A obeys...

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

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

- It's easy to prove that

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



Another fact about Multivalued Random Variables:

- Using the axioms of probability...
 $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- And assuming that A obeys...
 $P(A = v_i \wedge A = v_j) = 0$ if $i \neq j$
 $P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$
- It's easy to prove that

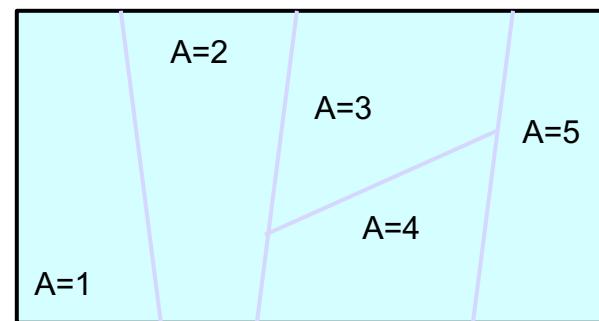
$$P(B \wedge [A = v_1 \vee A = v_2 \vee A = v_i]) = \sum_{j=1} P(B \wedge A = v_j)$$

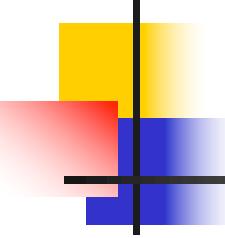
- And thus we can prove

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

Elementary Probability in Pictures

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





Independent Events

- Definition: two events A and B are *independent* if $\Pr(A \text{ and } B) = \Pr(A) * \Pr(B)$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Conditional Probability

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

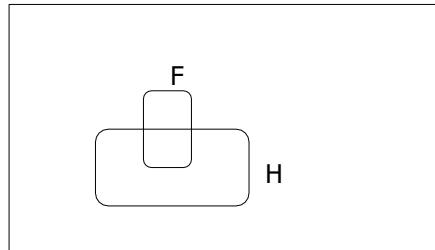
H = "Have a headache"

F = "Coming down with Flu"

$$P(H) = 1/10$$

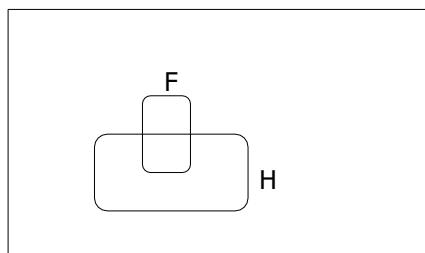
$$P(F) = 1/40$$

$$P(H|F) = 1/2$$



"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."

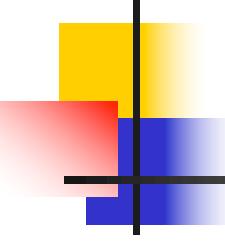
Conditional Probability



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

$$\begin{aligned}P(H) &= 1/10 \\ P(F) &= 1/40 \\ P(H|F) &= 1/2\end{aligned}$$

$$\begin{aligned}P(H | F) &= \frac{\text{Fraction of flu-inflicted worlds in which you have a headache}}{\text{#worlds with flu}} \\ &= \frac{\text{#worlds with flu and headache}}{\text{#worlds with flu}} \\ &= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}} \\ &= \frac{P(H \wedge F)}{P(F)}\end{aligned}$$



Definition of Conditional Probability

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

Corollary: The Chain Rule

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

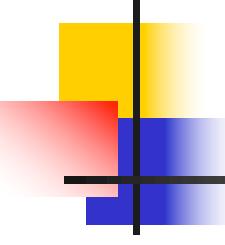
What we just did...

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

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**



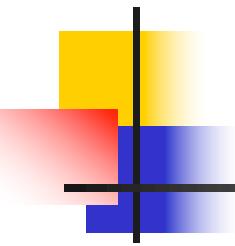


More General Forms of Bayes Rule

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

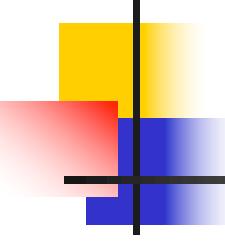
$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

$$P(A = a_1 | B) = \frac{P(B | A = a_1)P(A = a_1)}{\sum_i P(B | A = a_i)P(A = a_i)}$$



More General Forms of Bayes Rule

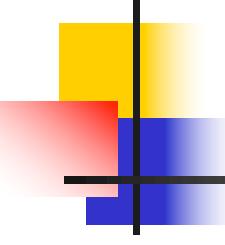
$$P(A=v_i|B) = \frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^{n_A} P(B|A=v_k)P(A=v_k)}$$



Useful Easy-to-prove facts

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

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



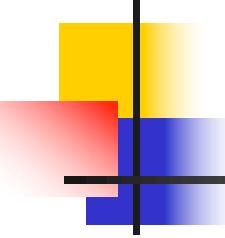
Useful Easy-to-prove facts

Other Forms of Bayes Rule

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

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

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



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

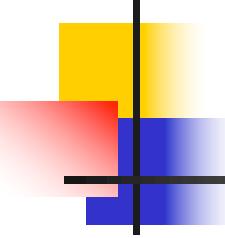
Assume:

$$P(A) = 0.05$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.20$$

what is $P(\text{flu} | \text{cough}) = P(A|B)$?

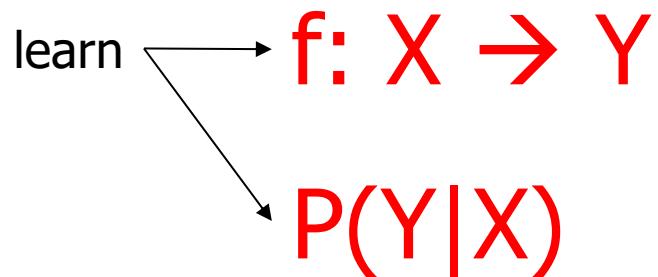


You Should Know

- 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



what does all this have to do
with function approximation?

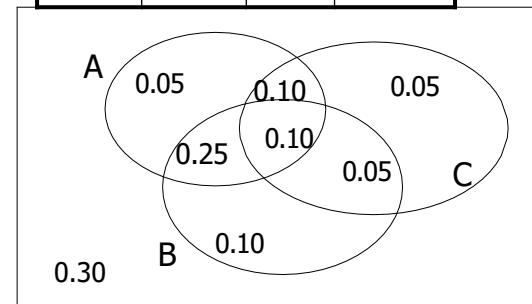


The Joint Distribution

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



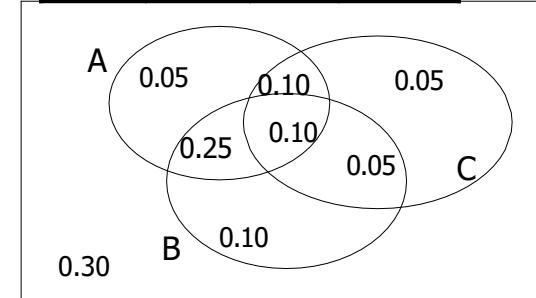
The Joint Distribution

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).

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



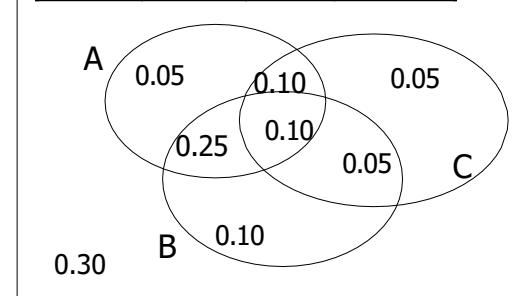
The Joint Distribution

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.

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



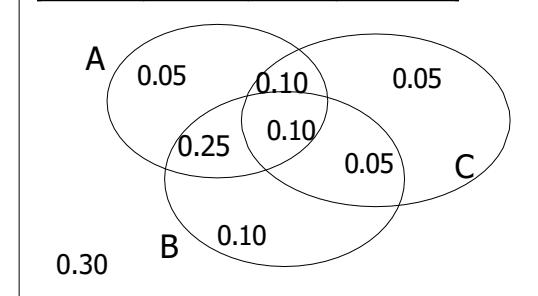
The Joint Distribution

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



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	?

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

Then 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

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

Once 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

$$P(\text{Poor Male}) = 0.4654$$

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

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

$$P(\text{Poor}) = 0.7604$$

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

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

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

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

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

$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$

Inference with the Joint

Learning and the Joint Distribution

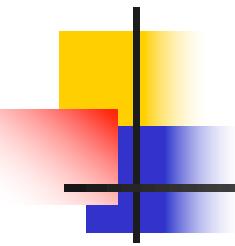
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function $f: \langle G, H \rangle \rightarrow W$

Equivalently, $P(W | G, H)$

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

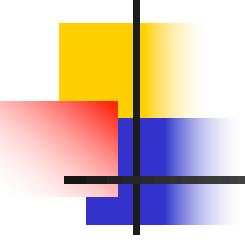
e.g., $P(W=\text{rich} | G = \text{female}, H = 40.5-) =$



sounds like the solution to
learning $F:X \rightarrow Y$,
or $P(Y|X)$

Are we done?

consider learning Joint Dist. with 100 attributes
of rows in this table?
of people on earth?
fraction of rows with 0 training examples?



Our first machine learning problem:

Parameter estimation: MLE, MAP

Estimating Probabilities



Flipping a Coin

I have a coin, if I flip it, what's the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:



The estimated probability is: $3/5$ “Frequency of heads”

Flipping a Coin

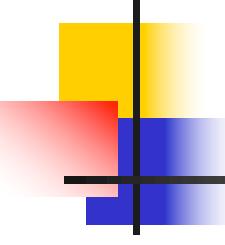


The estimated probability is: **3/5** “Frequency of heads”

Questions:

- (1) Why frequency of heads???**
- (2) How good is this estimation???**
- (3) Why is this a machine learning problem???**

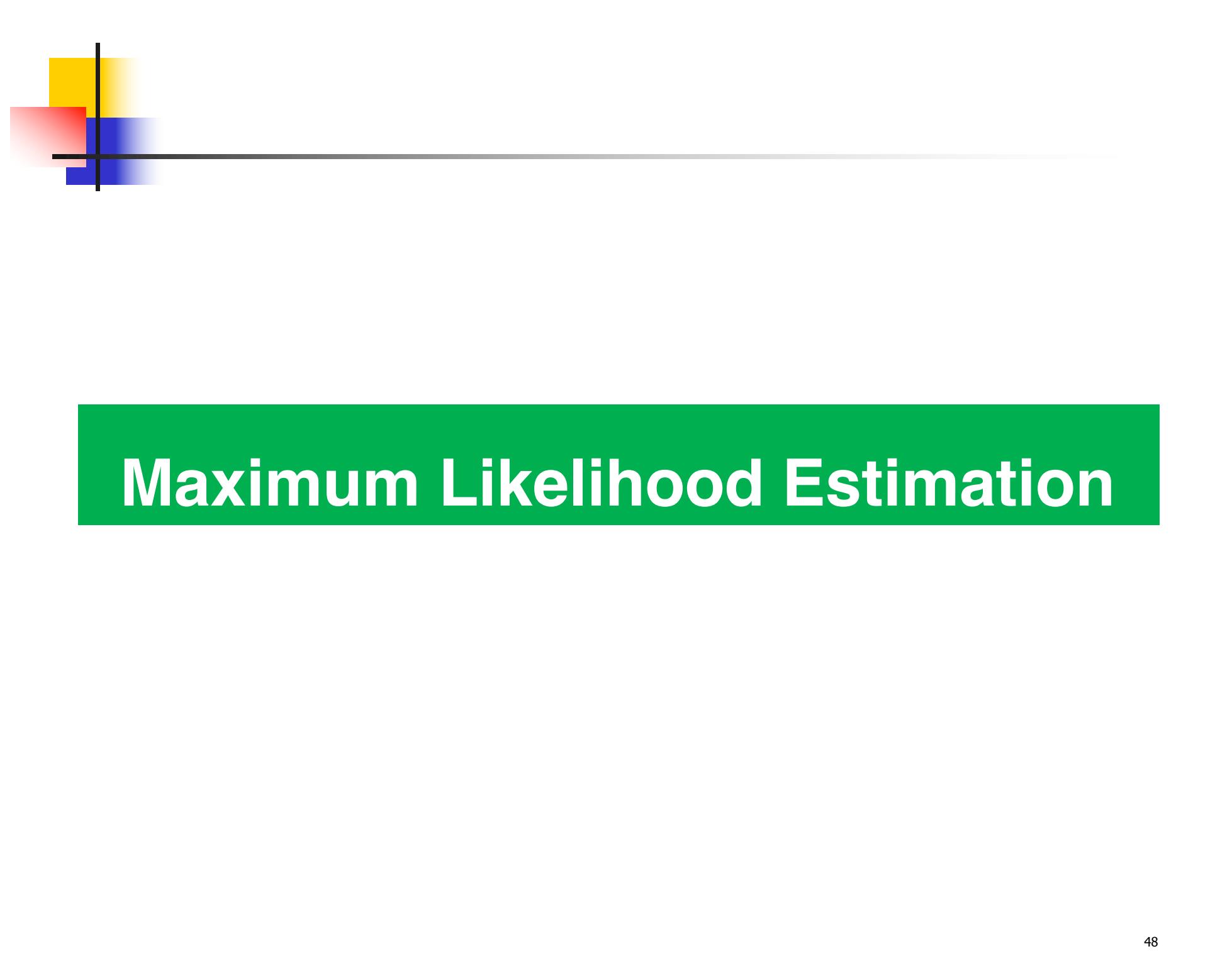
We are going to answer these questions



Question (1)

Why frequency of heads???

- Frequency of heads is exactly the ***maximum likelihood estimator*** for this problem
- MLE has nice properties
(interpretation, statistical guarantees, simple)



Maximum Likelihood Estimation

MLE for Bernoulli distribution

Data, $D =$



$$D = \{X_i\}_{i=1}^n, \quad X_i \in \{\text{H}, \text{T}\}$$

$$P(\text{Heads}) = \theta, \quad P(\text{Tails}) = 1-\theta$$

Flips are i.i.d.:

- **Independent** events
- **Identically distributed** according to Bernoulli distribution

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D \mid \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i \mid \theta) \quad \text{Independent draws} \\ &= \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta) \quad \text{Identically distributed} \\ &= \arg \max_{\theta} \underbrace{\theta^{\alpha_H} (1 - \theta)^{\alpha_T}}_{J(\theta)}\end{aligned}$$

Maximum Likelihood Estimation

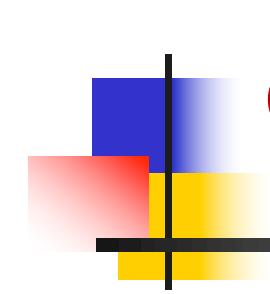
MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D \mid \theta) \\ &= \arg \max_{\theta} \underbrace{\theta^{\alpha_H} (1 - \theta)^{\alpha_T}}_{J(\theta)}\end{aligned}$$

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= \alpha_H \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T - 1} \Big|_{\theta=\hat{\theta}_{MLE}} = 0 \\ \alpha_H (1 - \theta) - \alpha_T \theta &\Big|_{\theta=\hat{\theta}_{MLE}} = 0\end{aligned}$$

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

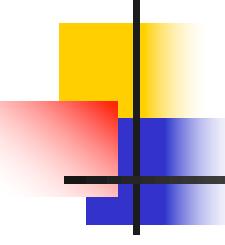
That's exactly the "Frequency of heads"



Question (2)

How good is this MLE estimation???

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



How many flips do I need?

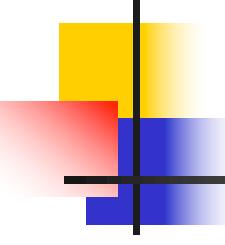
I flipped the coins 5 times: 3 heads, 2 tails

$$\hat{\theta}_{MLE} = \frac{3}{5}$$

What if I flipped 30 heads and 20 tails?

$$\hat{\theta}_{MLE} = \frac{30}{50}$$

- **Which estimator should we trust more?**
- **The more the merrier???**

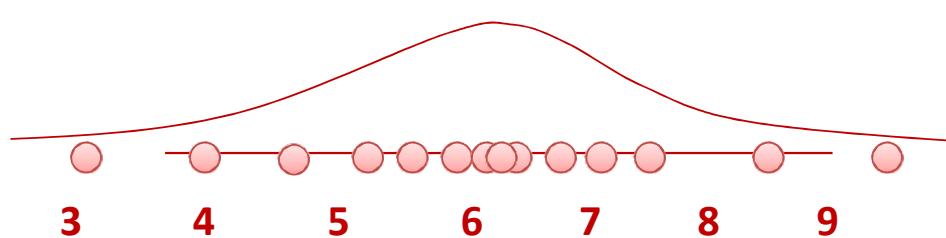


Question (3)

Why is this a machine learning problem???

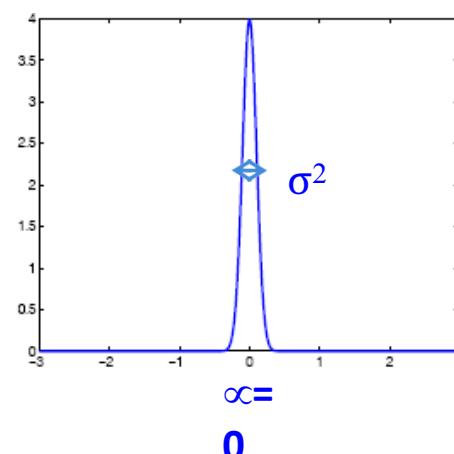
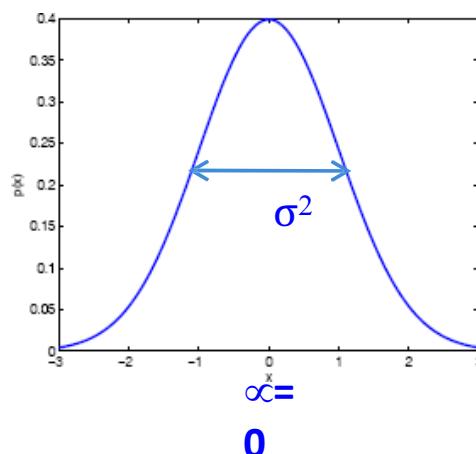
- improve their **performance** (accuracy of the predicted prob.)
- at some **task** (predicting the probability of heads)
- with **experience** (the more coins we flip the better we are)

What about continuous features?



Let us try Gaussians...

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma)$$



MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i | \theta) \quad \text{Independent draws} \\ &= \arg \max_{\theta} \prod_{i=1}^n \frac{1}{2\sigma^2} e^{-(X_i - \mu)^2 / 2\sigma^2} \quad \text{Identically distributed} \\ &= \arg \max_{\theta=(\mu, \sigma^2)} \underbrace{\frac{1}{2\sigma^2} e^{-\sum_{i=1}^n (X_i - \mu)^2 / 2\sigma^2}}_{J(\theta)}\end{aligned}$$

MLE for Gaussian mean and variance

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator: $\hat{\sigma}_{unbiased}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$

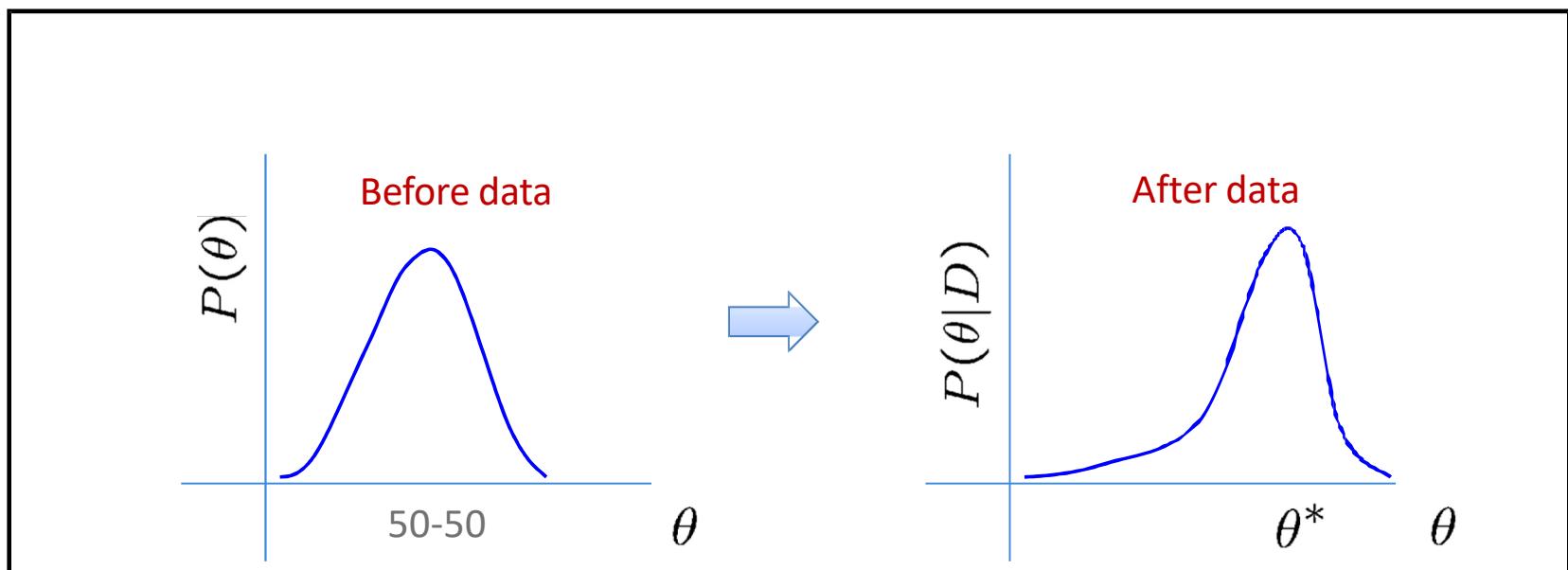
$$E[\hat{\sigma}_{MLE}^2] \neq \sigma^2 \quad E[\hat{\sigma}_{unbiased}^2] = \sigma^2$$



What about prior knowledge? (MAP Estimation)

What about prior knowledge?

- We know the coin is “close” to 50-50. What can we do now?
- **The Bayesian way...**
- Rather than estimating a single θ , we obtain a distribution over possible values of θ



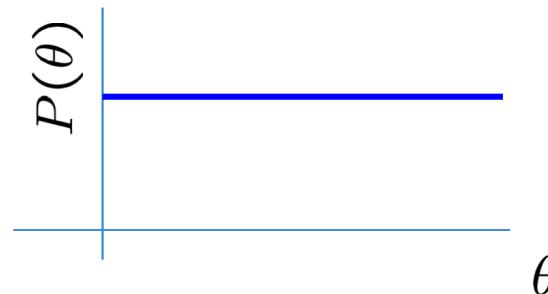
Prior distribution

What prior? What distribution do we want for a prior?

- Represents expert knowledge (**philosophical approach**)
- Simple posterior form (**engineer's approach**)

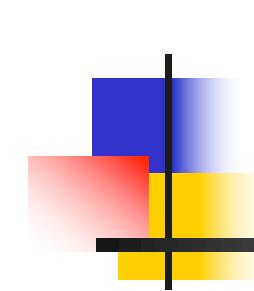
Uninformative priors:

- Uniform distribution



Conjugate priors:

- Closed-form representation of posterior
- $P(\theta)$ and $P(\theta|ID)$ have the same form

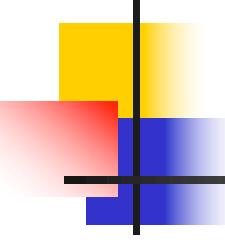


In order to proceed we will need:

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



Chain Rule & Bayes Rule

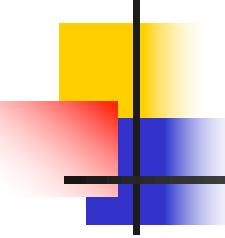
Chain rule:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Bayes rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes rule is important for reverse conditioning.



Bayesian Learning

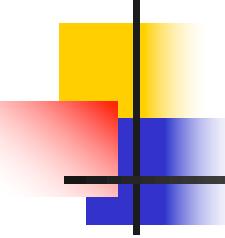
- Use Bayes rule:

$$P(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$$

posterior likelihood prior



MLE vs. MAP

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

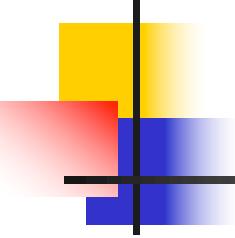
$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta|D) \\ &= \arg \max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?



MAP estimation for Binomial distribution

Coin flip problem: Likelihood is Binomial

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

If the prior is Beta distribution,

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

⇒ posterior is Beta distribution

Beta function: $B(x, y) = \int_0^1 t^{x-1} (1 - t)^{y-1} dt$

MAP estimation for Binomial distribution

Likelihood is Binomial: $P(\mathcal{D} | \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$

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

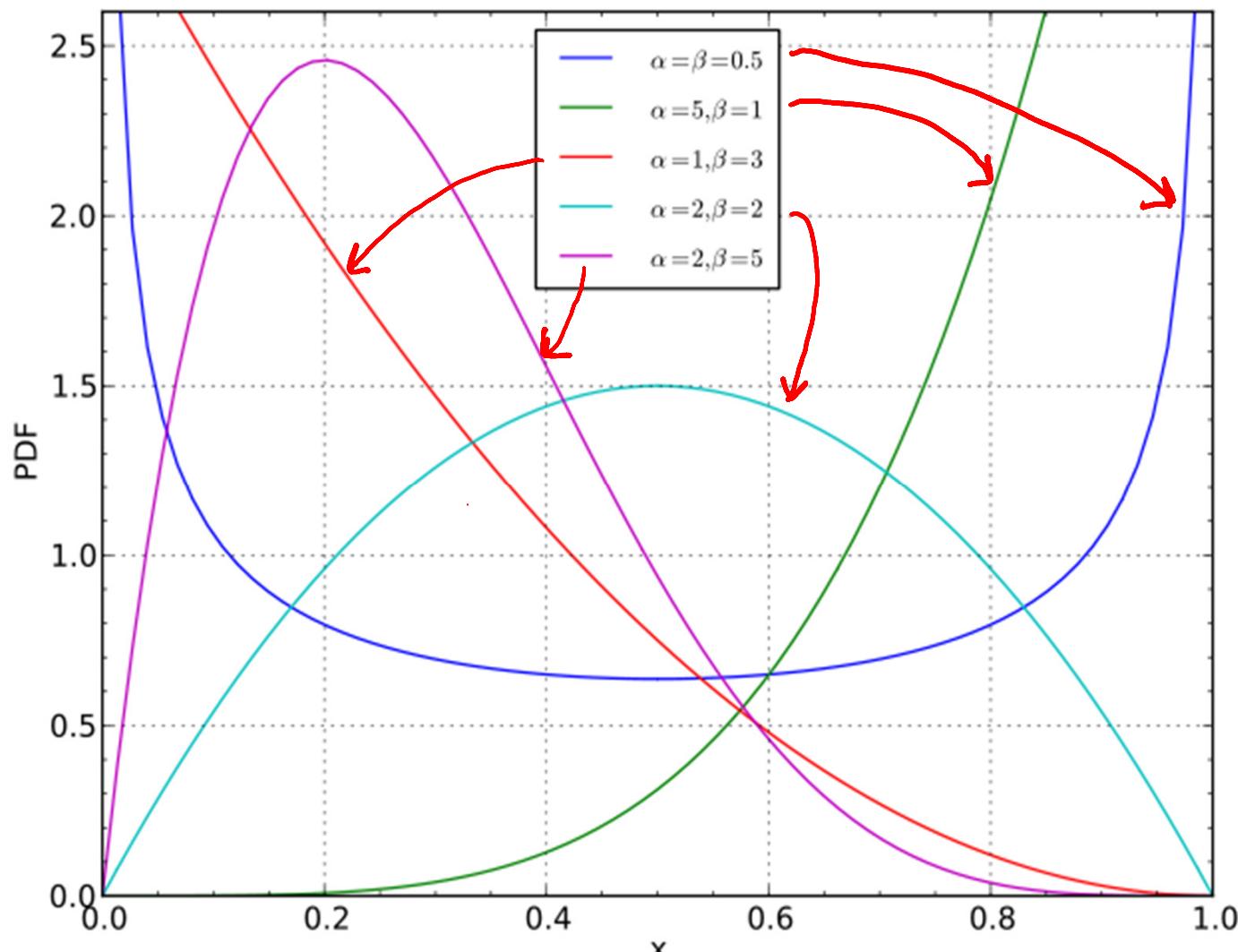
⇒ posterior is Beta distribution

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

$P(\theta)$ and $P(\theta|D)$ have the same form! [Conjugate prior]

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} P(D | \theta)P(\theta) \\ &= \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}\end{aligned}$$

Beta distribution

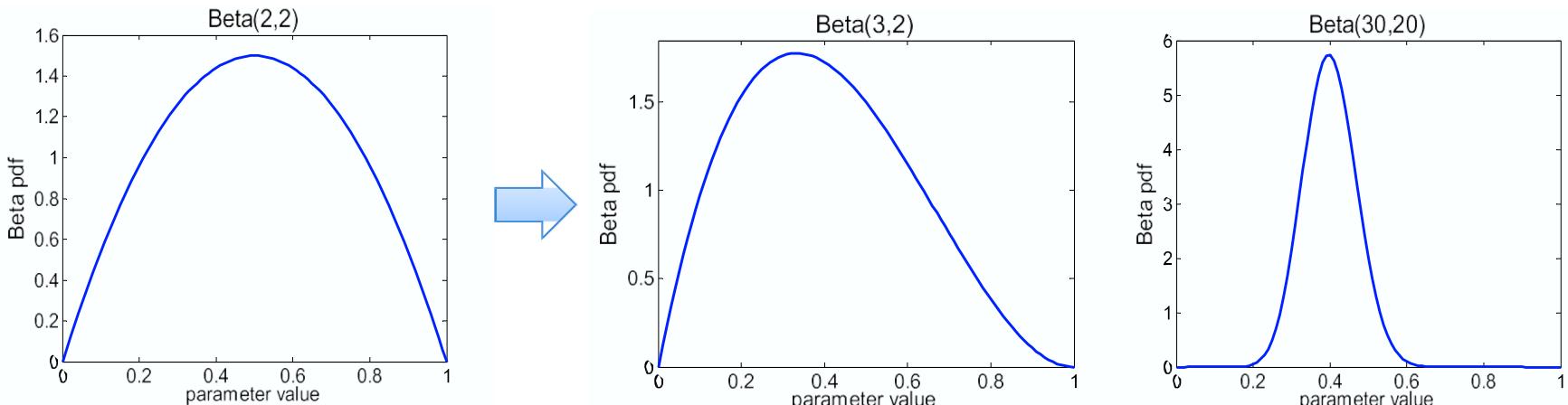


More concentrated as values of α , β increase

Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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



As $n = \alpha_H + \alpha_T$
increases

As we get more samples, effect of prior is “washed out”

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

From Binomial to Multinomial

Example: Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim \text{Multinomial}(\theta = \{\theta_1, \theta_2, \dots, \theta_k\})$



$$P(\mathcal{D} | \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \dots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

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

Then posterior is Dirichlet distribution

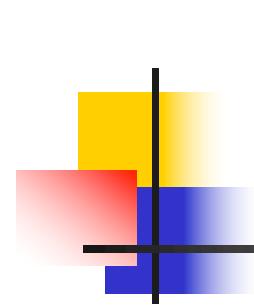
$$P(\theta | D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k + \alpha_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^k (\alpha_j + \beta_j - 1)}$$

For Multinomial, conjugate prior is Dirichlet distribution.

http://en.wikipedia.org/wiki/Dirichlet_distribution



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, ...