Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 18, 2011

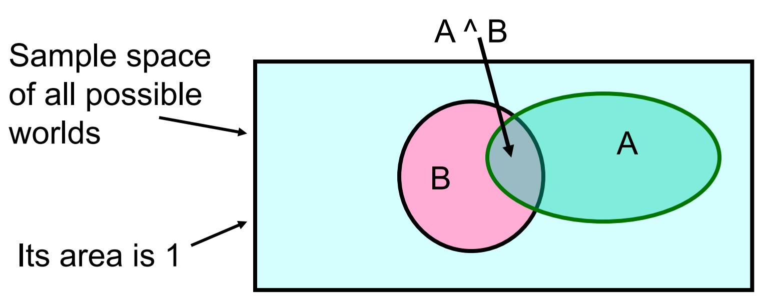
Today:
- Bayes Rule
- Estimating parameters
  - maximum likelihood
  - max a posteriori

many of these slides are derived from William Cohen, Andrew Moore, Aarti Singh, Eric Xing, Carlos Guestrin. - Thanks!

Readings:

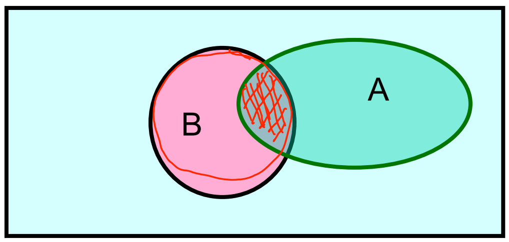
Probability review
- Bishop Ch. 1 thru 1.2.3
- Bishop, Ch. 2 thru 2.2
- Andrew Moore's online tutorial

Visualizing Probabilities



Definition of Conditional Probability

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



Definition of Conditional Probability

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

implies

Corollary: The Chain Rule

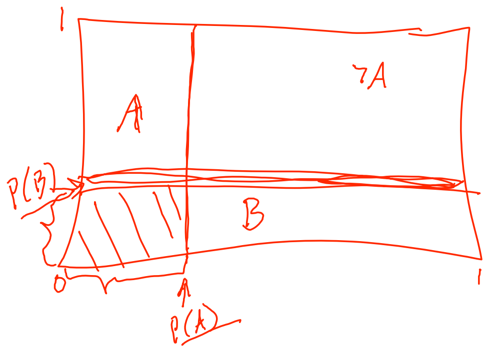
$$P(A \wedge B) = P(A|B)\, P(B)$$

$$P(C \wedge A \wedge B) = P(C|A \wedge B)\, \overbrace{P(A|B)\, P(B)}^{P(A \wedge B)}$$

Independent Events

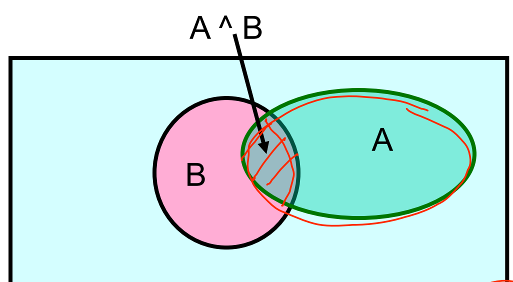
$P(A|B) = \frac{P(A \wedge B)}{P(B)} =$

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



Bayes Rule

- let's write 2 expressions for $\boxed{P(A \wedge B)}$



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

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

Bayes Rule!

evidence Prior

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

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

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

Applying Bayes Rule

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

flu coughed

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

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

$P(\neg A) = 1 - P(A) = .95$

$$\frac{.8 \; .05}{.8 \; .05 + 0.2 \; .95} = \frac{.04}{.04 + .19}$$

$$\frac{.04}{.23} \approx .18$$

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

what does all this have to do with function approximation?

learn → $\boxed{f: X \to Y}$

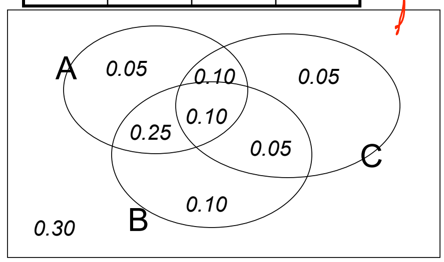
→ $\boxed{P(Y|X)}$

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

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



[A. Moore]

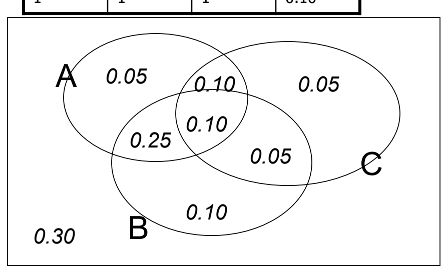
The Joint Distribution

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



[A. Moore]

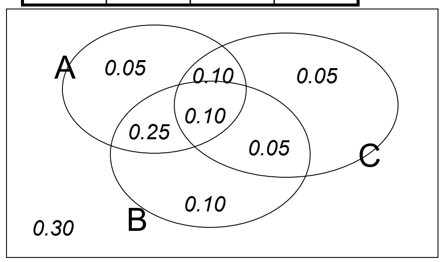
The Joint Distribution

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



[A. Moore]

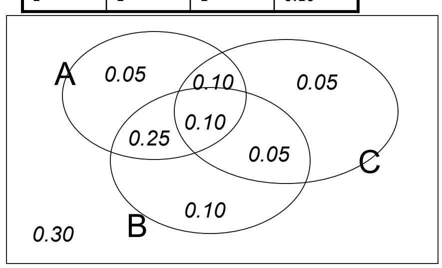
The Joint Distribution

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



[A. Moore]

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

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

[A. Moore]

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

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

[A. Moore]

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

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

[A. Moore]

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

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

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

[A. Moore]

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

[A. Moore]

sounds like the solution to
learning F: X →Y,
or P(Y | X).

Are we done?

No.

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 with the nail up?
  - You say: Please flip it a few times:

  - You say: The probability is: $\boxed{0.6}$?
  - **He says: Why???**
  - You say: Because…

[C. Guestrin]

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 outcomes (Binomial)

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

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

[C. Guestrin]

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's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

data likelihood

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

log likelihood

[C. Guestrin]

Maximum Likelihood Estimate for Θ

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

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

[C. Guestrin]

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

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

$$\frac{\partial}{\partial \theta} \quad = \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T} = \alpha_H \ln \theta + \alpha^T \ln(1-\theta)$$

$$\alpha_H \frac{\partial}{\partial \theta} \ln \theta + \alpha^T \frac{\partial}{\partial \theta} \ln(1-\theta)$$

$$\boxed{\frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}}$$

$$\alpha_H \frac{1}{\theta} + \alpha^T \frac{\partial \quad (1-\theta)}{\partial \theta} \qquad \boxed{\frac{\partial \ln(1-\theta)}{\partial (1-\theta)}}$$

$$\boxed{\alpha_H \frac{1}{\theta} + \alpha^T(-1)\frac{1}{1-\theta} = 0}$$

$$\vdots$$

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

[C. Guestrin]

How many flips do I need?

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

$$\frac{3}{3+2} = .6 = \text{MLE } \theta$$

$$\frac{300}{300+200} = .6 = \text{MLE } \theta$$

[C. Guestrin]

Bayesian Learning

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

- Use Bayes rule:

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

Not dep. on θ

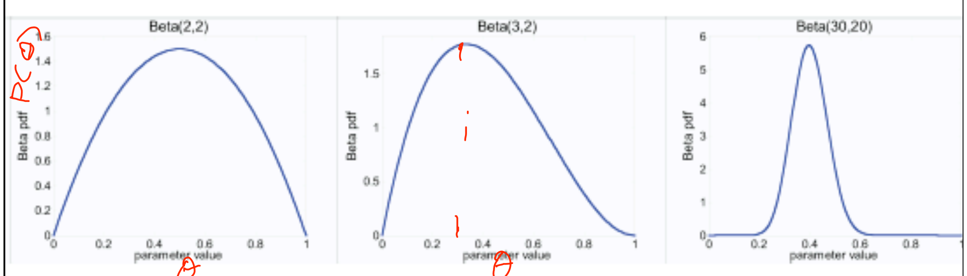
- Or equivalently: $= \arg\max_\theta P(\mathcal{D}\mid\theta)P(\theta)$

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

[C. Guestrin]

Beta prior distribution – P(θ)

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



[C. Guestrin]

Beta prior distribution – P(θ)

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

2 2

- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $\underset{\theta}{\arg\max}\, P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

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

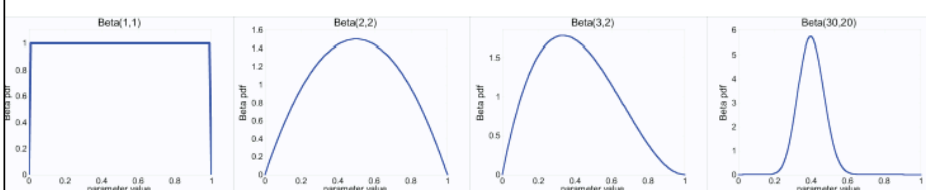
$$\underset{\theta}{\arg\max} \frac{\theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}}{Z(\ )} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

[C. Guestrin]

Posterior distribution

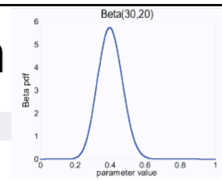
- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

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



[C. Guestrin]

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:

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

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

[C. Guestrin]

Conjugate priors

- P(θ) and P(θ|D) have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

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

For Binomial, conjugate prior is Beta distribution.

[A. Singh]

Conjugate priors

- $P(\theta)$ and $P(\theta|D)$ have the same form

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

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

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

If prior is Dirichlet distribution,

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

Then posterior is Dirichlet distribution

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

For Multinomial, conjugate prior is Dirichlet distribution.

[A. Singh]

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

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

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

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

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

Dirichlet 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

$$P(\theta_1, \theta_2, ...\theta_K) = \frac{1}{B(\alpha)} \prod_{i}^{K} \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

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 posteriori estimates
  – distributions – binomial, Beta, Dirichlet, …
  – conjugate priors

Extra slides

Expected values

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

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

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

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

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

or   X=gender, Y=leftHanded

Remember: $E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$

Example: Bernoulli model

- Data:
  - We observed N ***iid*** coin tossing: D={1, 0, 1, …, 0}
- Representation:

  Binary r.v: $x_n = \{0,1\}$
- Model: $P(x) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \Rightarrow P(x) = \theta^x (1-\theta)^{1-x}$
- How to write the likelihood of a single observation x_i?

  $$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$
- The likelihood of dataset $D=\{x_1, \dots, x_N\}$:

$$P(x_1, x_2, \dots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}} (1-\theta)^{\#\text{tails}}$$