Foundations

COMP3314 — Lecture 2

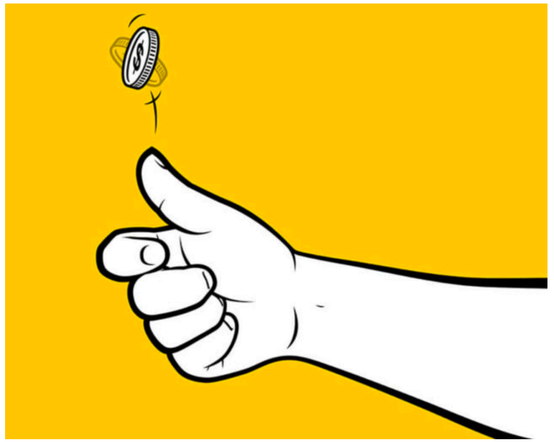
Lingpeng Kong
Department of Computer Science, The University of Hong Kong

Based on: Probabilistic Machine Learning by Kevin Murphy
Slides from: Saw Shier Nee with special thanks!

What is probability?

P(head) = 50%

Frequentist



events (repeated trails)

Bayesian

uncertainty (ignorance)

Probability

Joint probability:

$$\Pr(A \wedge B) = \Pr(A, B)$$

Probability of a union of two events:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B)$$

Conditional probability:

$$\Pr(B|A) \triangleq \frac{\Pr(A, B)}{\Pr(A)}$$

If A and B are independent events:

$$\Pr(A, B) = \Pr(A) \Pr(B)$$

If the events are mutually exclusive:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B)$$

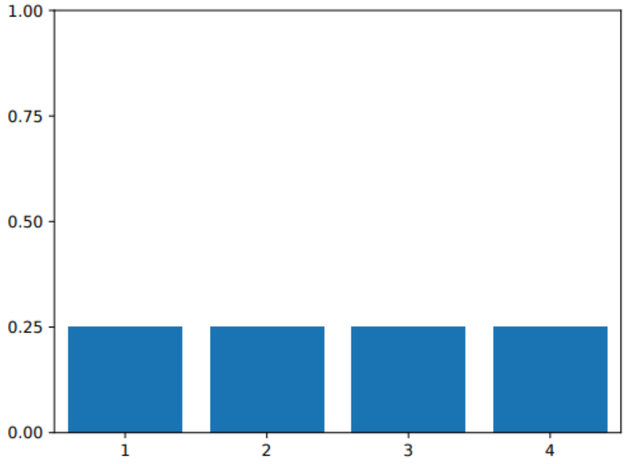
Conditional independence:

$$\Pr(A, B|C) = \Pr(A|C) \Pr(B|C)$$

Random Variables

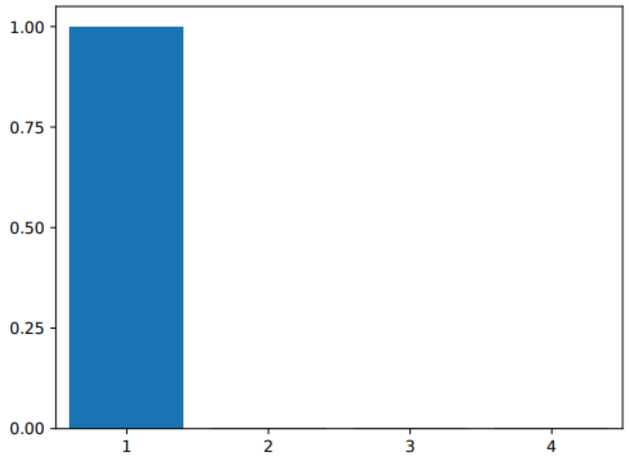
X represents some unknown quantity of interest

We call it a **random variable** if the value of X is unknown and/or could change.



(a)

uniform distribution

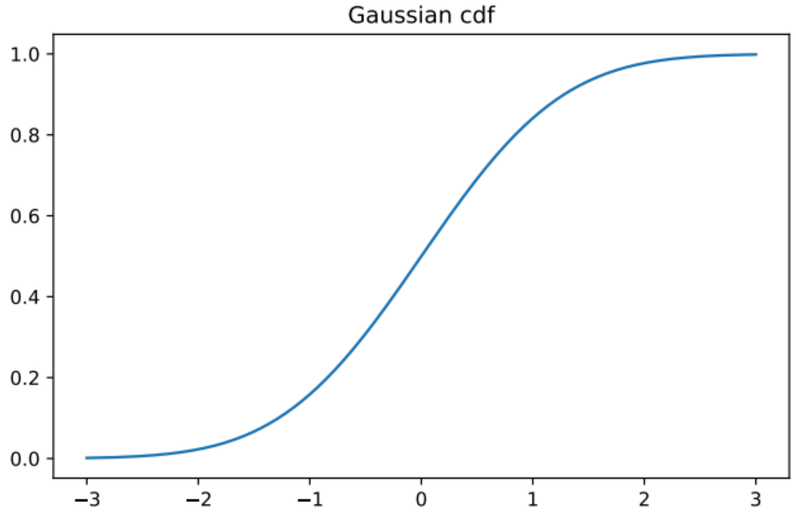


(b)

degenerate distribution
(delta function)

probability mass function (pmf): $p(x) \triangleq \Pr(X = x)$

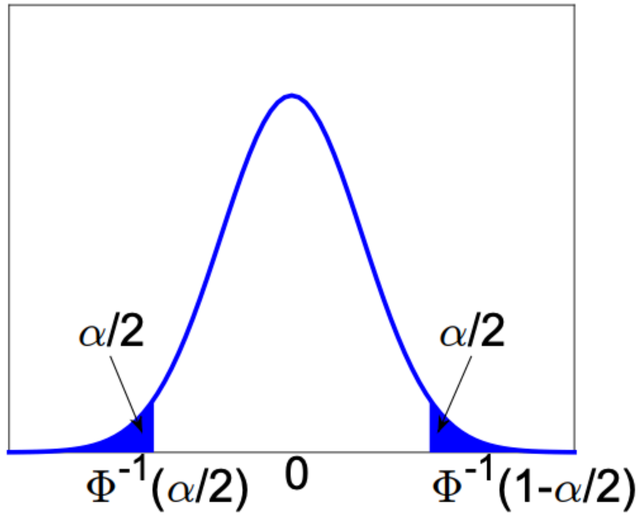
Continuous Random Variables



(a)

cumulative distribution function (cdf)

$$P(x) \triangleq \Pr(X \leq x)$$



(b)

probability density function (pdf)

$$p(x) \triangleq \frac{d}{dx} P(x)$$

$$\Pr(a < X \leq b) = \int_a^b p(x)dx = P(b) - P(a)$$

$$\Pr(x \leq X \leq x + dx) \approx p(x)dx$$

Sets of Related Random Variables

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	0.2	0.3
$X=1$	0.3	0.2

marginal distribution

$$p(X=x)=\sum_y p(X=x,Y=y)$$

conditional distribution

$$p(Y=y|X=x)=\frac{p(X=x,Y=y)}{p(X=x)}$$

$$p(x,y)=p(x)p(y|x)$$

chain rule

$$p(\boldsymbol{x}_{1:D})=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3)\ldots p(x_D|\boldsymbol{x}_{1:D-1})$$

Moments of a Distribution

Mean of a distribution

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\, p(x) dx$$

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\, p(x)$$

Linearity of expectation

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Variance of a distribution

$$\begin{aligned}\mathbb{V}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] &= \int (x-\mu)^2 p(x) dx \\ &= \int x^2 p(x) dx + \mu^2 \int p(x) dx - 2\mu \int x p(x) dx = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

$$\mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2$$ assuming some random variable with mean = 0

$$\mathbb{V}[aX + b] = a^2 \mathbb{V}[X]$$

Moments of a Distribution

Mode of a distribution

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x})$$

conditional moments

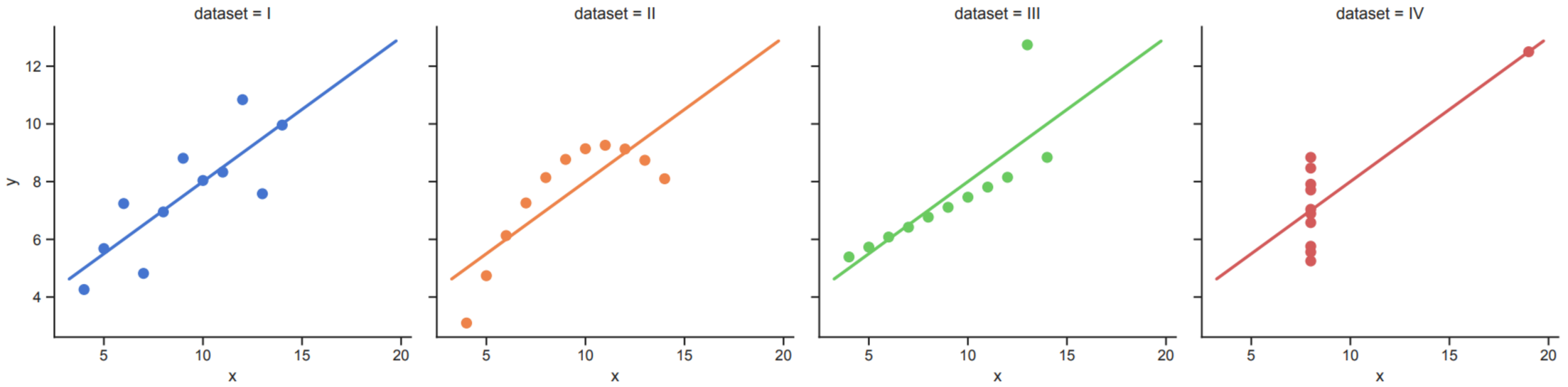
$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$$

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}_Y\left[\sum_x x\, p(X = x|Y)\right]$$

$$= \sum_y \left[\sum_x x\, p(X = x|Y)\right] p(Y = y) = \sum_{x,y} x p(X = x, Y = y) = \mathbb{E}[X]$$

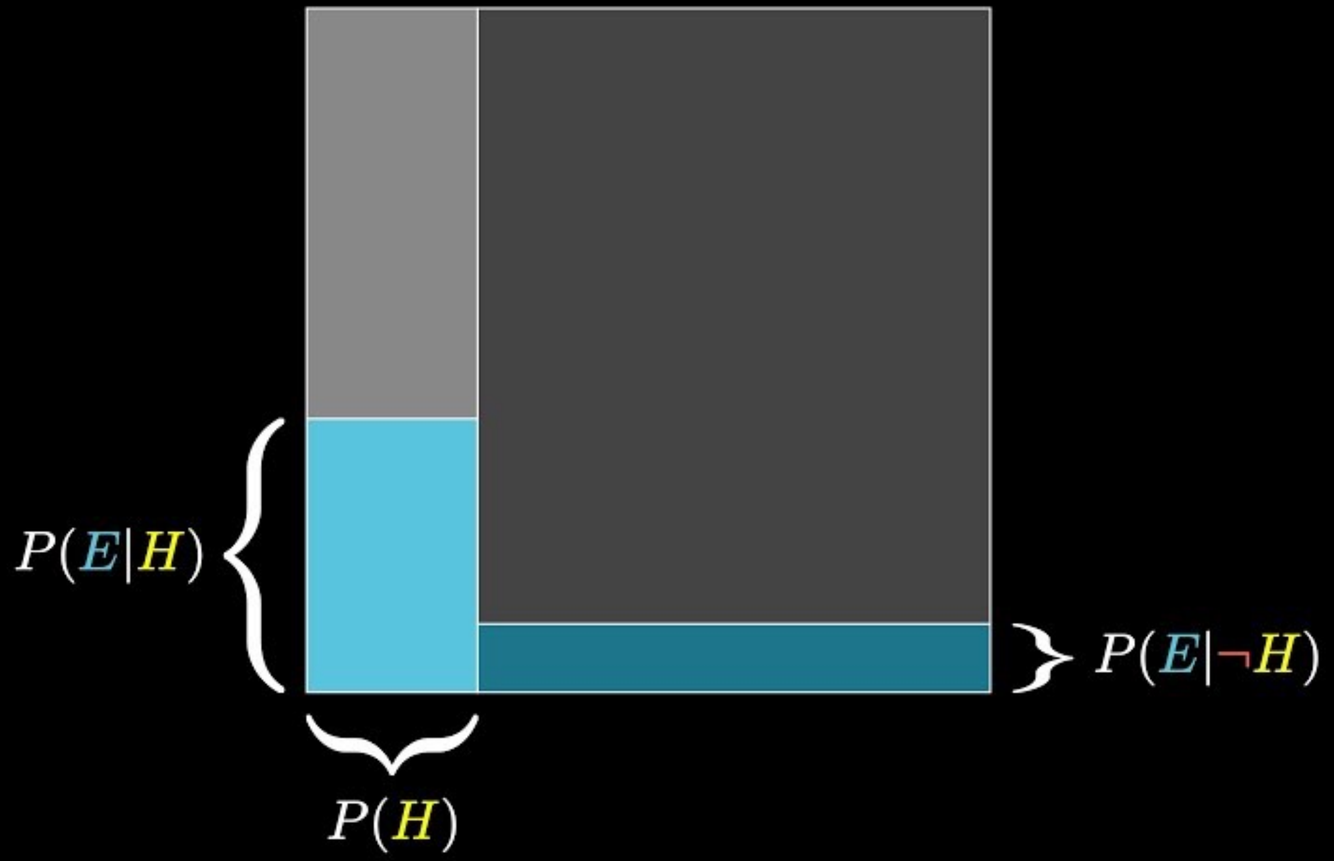
law of total (iterated) expectation

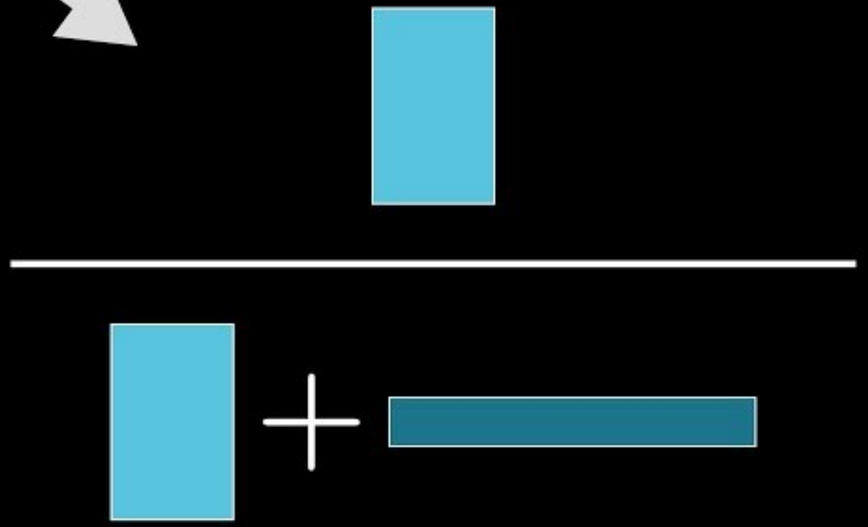
Limitations of Summary Statistics



Anscombe's Quartet: E[x] = 9, V[x] = 11, E[y] = 7.50, V[y]= 4.125

This is Bayes' rule





Bayes' rule

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

prior distribution — what we know about possible values of H before we see any data

$$p(H)$$

observation distribution — possible outcomes Y we expect to see if H =h

$$p(Y|H = h)$$

likelihood — evaluate the observation distribution at a point corresponding to the actual observations, y

$$p(Y = y|H = h)$$

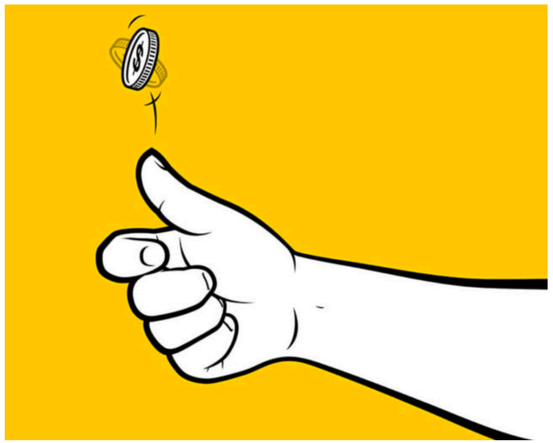
marginal likelihood

$$p(Y = y) = \sum_{h' \in \mathcal{H}} p(H = h')p(Y = y|H = h') = \sum_{h' \in \mathcal{H}} p(H = h', Y = y)$$

posterior — our new belief state about the possible value of H

$$p(H = h|Y = y)$$

Bernoulli Distribution

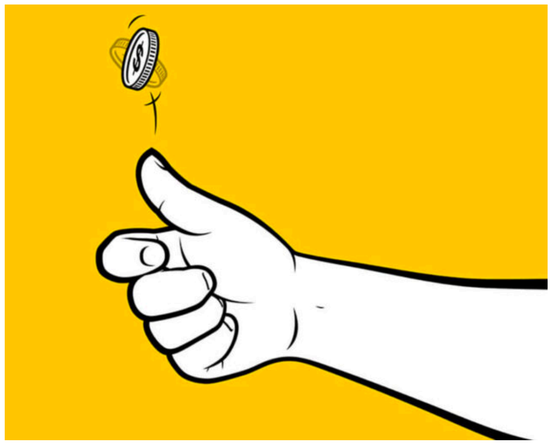


$$Y \sim \text{Ber}(\theta)$$

$$\text{Ber}(y|\theta) = \begin{cases} 1-\theta & \text{if } y = 0 \\ \theta & \text{if } y = 1 \end{cases}$$

$$\text{Ber}(y|\theta) \triangleq \theta^y (1-\theta)^{1-y}$$

Binomial Distribution



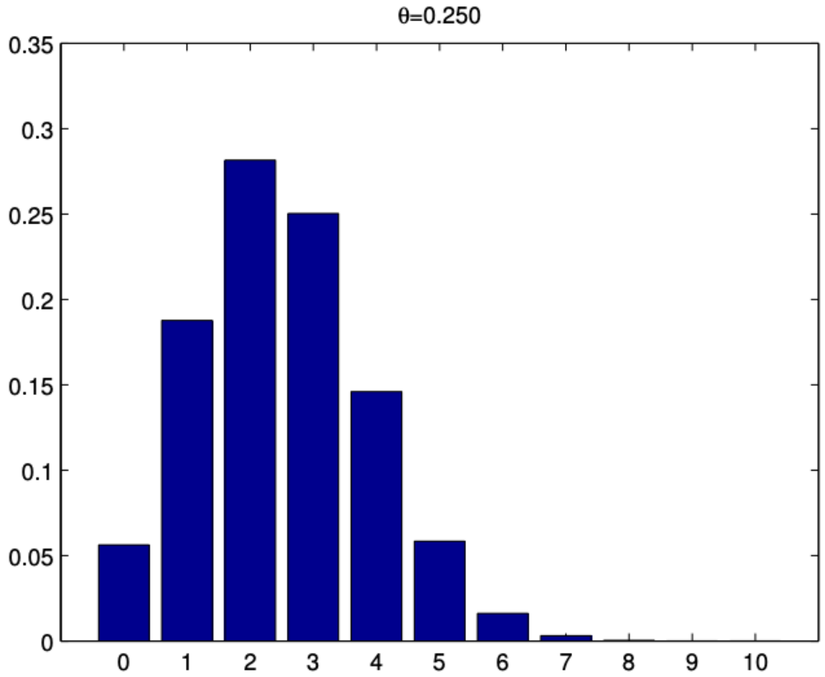
N = 5 (N Bernoulli trials), s = 4 (total number of heads)

$$\text{Bin}(s|N,\theta) \triangleq \binom{N}{s}\theta^s(1-\theta)^{N-s}$$

$$\binom{N}{k} \triangleq \frac{N!}{(N-k)!k!}$$

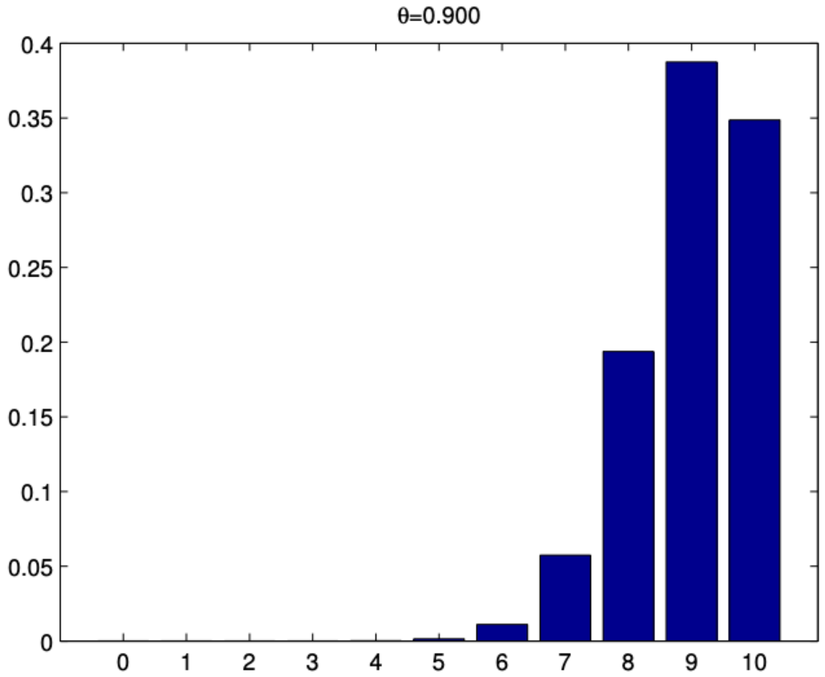
Binomial Distribution

$$\text{Bin}(s|N,\theta) \triangleq \binom{N}{s}\theta^s(1-\theta)^{N-s}$$



(a)

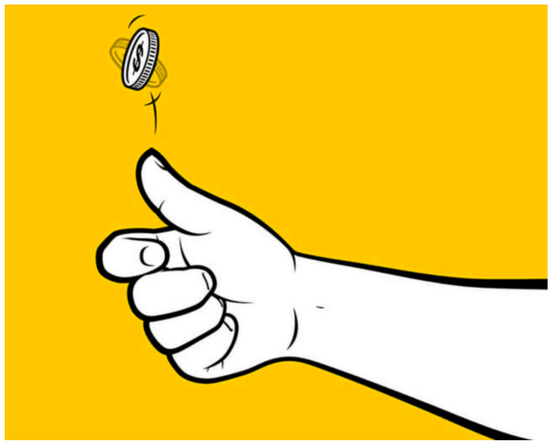
$N = 10, \theta = 0.25$



(b)

$N = 10, \theta = 0.9$

A Simple Question?



What's the probability of head up? P(Heads)

0.8?

A Simple Question?

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

$$D = \{\text{head}, \text{head}, \text{head}, \text{head}, \text{tail}\}$$

α_H total number of heads α_T total number of tails

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

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

Set derivative to zero

$$\frac{\partial \log P(D \mid \theta)}{\partial \theta} = 0$$

Maximum Likelihood Estimation

Set derivative to zero

$$\frac{\partial \log P(D \mid \theta)}{\partial \theta} = 0$$

$$\frac{\partial \log \theta^{\alpha_H}(1-\theta)^{\alpha_T}}{\partial \theta} = 0$$

$$\hat{\theta}_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4}{4+1} = 0.8$$

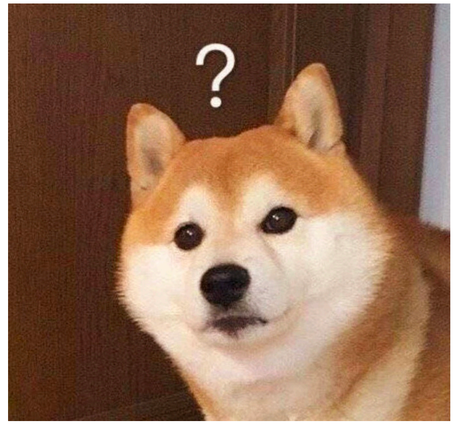
Maximum Likelihood Estimation

If you flip the coin 5 times, get 4 times head

$$\hat{\theta}_{\mathrm{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4}{4+1} = 0.8$$

If you flip the coin 5000 times, get 4000 times head

$$\hat{\theta}_{\mathrm{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4000}{4000+1000} = 0.8$$



Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

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

Set derivative to zero

$$\frac{\partial \log P(D \mid \theta)}{\partial \theta} = 0$$

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

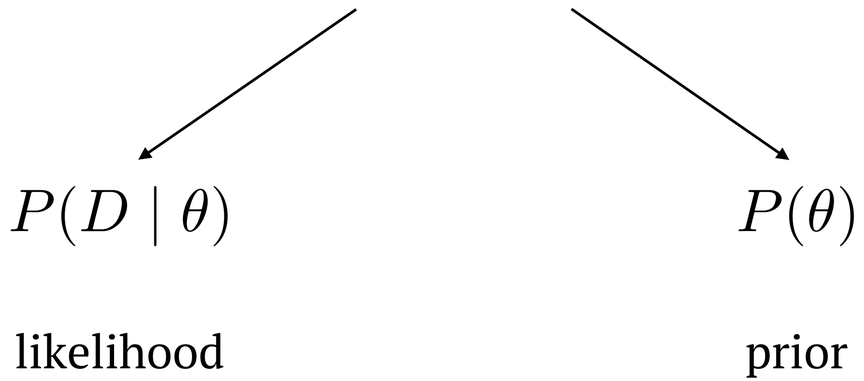
Bayesian Learning

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

posterior

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

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



$P(D \mid \theta)$ likelihood

$P(\theta)$ prior

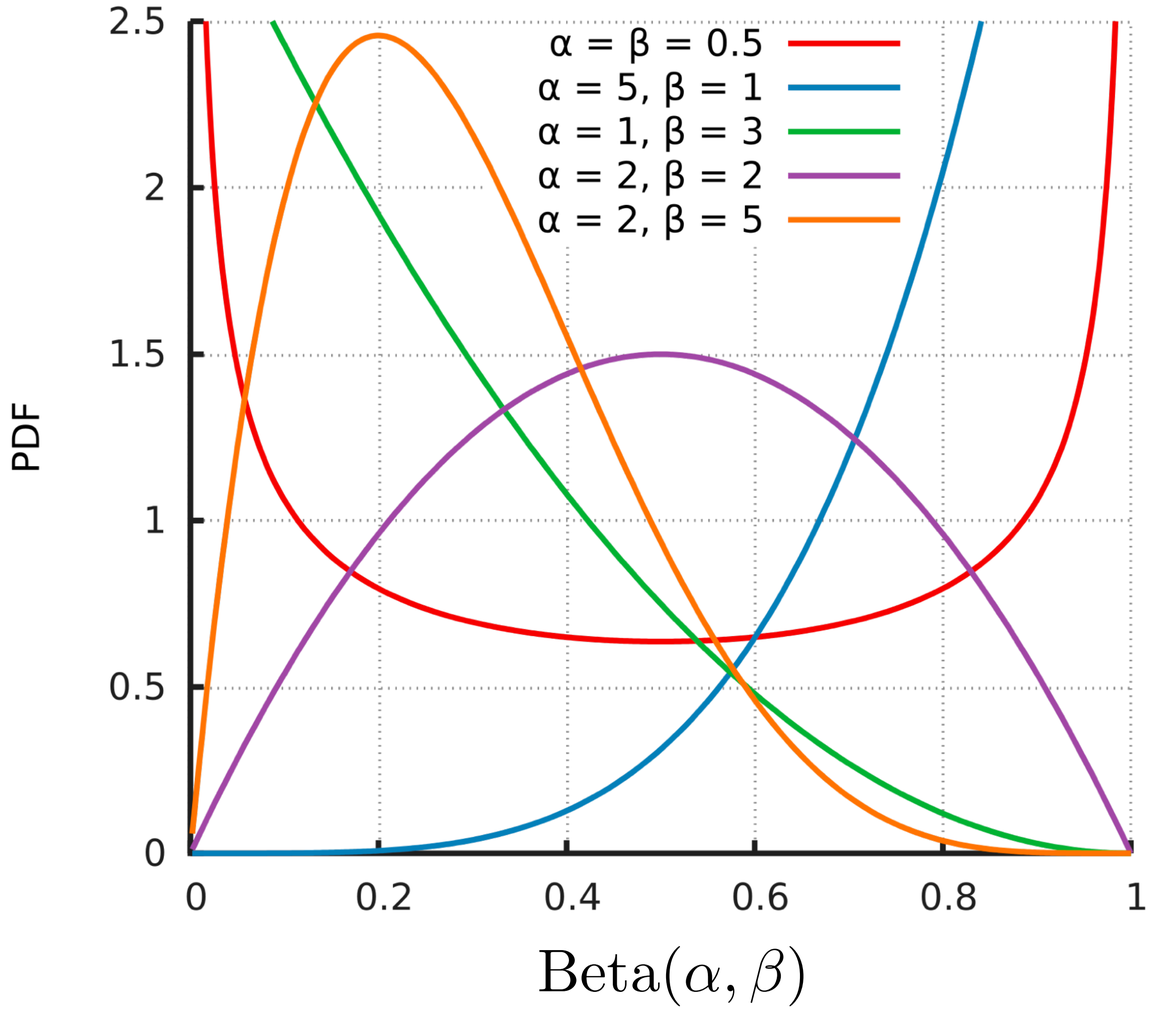
Bayesian Learning

prior distribution

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

Beta Distribution

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



Bayesian Learning

prior distribution

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

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

likelihood function

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

posterior distribution

$$P(\theta \mid D) \propto \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}$$

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

Estimating Parameters

MLE: Maximum Likelihood Estimate, choose θ that maximizes the probability of observed data

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

MAP: Maximum a Posteriori, choose θ that is most probable given prior probability and the data

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

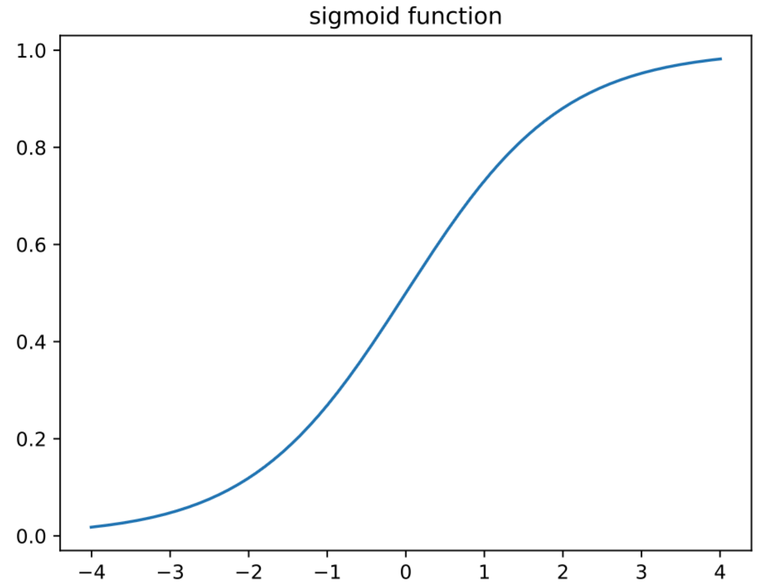
Sigmoid (logistic) function

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \text{Ber}(y|f(\boldsymbol{x}; \boldsymbol{\theta})) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad \text{Ber}(y|\theta) \triangleq \theta^y (1-\theta)^{1-y}$$

We want this to be an unconstrained function

sigmoid (logistic) function $\boldsymbol{\sigma}()$ helps:

$$\boldsymbol{\sigma}(a) \triangleq \frac{1}{1+e^{-a}} \qquad a = f(\boldsymbol{x}; \boldsymbol{\theta})$$



one type of “activation” in neural nets!

Categorical and Multinomial Distributions

categorical distribution

$$\mathcal{M}(y|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

$$\mathcal{M}(\boldsymbol{y}|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{y_c} \qquad C = 3 \quad (1,0,0)(0,1,0)(0,0,1)$$

one-hot vector

multinomial distribution

rolling a C-sided dice N times, **s** to be a vector that counts the number of times each face shows up

$$\mathcal{M}(\boldsymbol{s}|N,\boldsymbol{\theta}) \triangleq \binom{N}{s_1 \ldots s_C} \prod_{c=1}^{C} \theta_c^{s_c} \qquad \binom{N}{s_1 \ldots s_C} \triangleq \frac{N!}{s_1! s_2! \cdots s_C!}$$

Softmax function

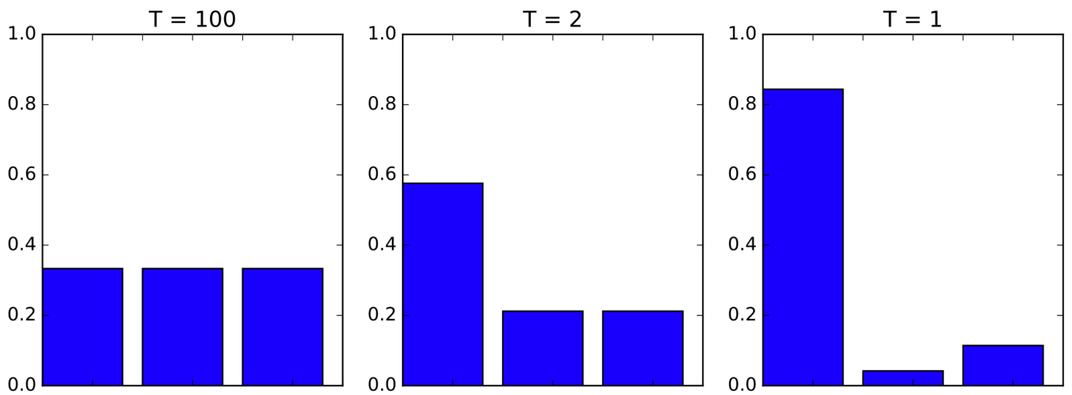
$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{M}(y|f(\boldsymbol{x}; \boldsymbol{\theta})) \qquad \mathcal{M}(y|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

We want this to be an unconstrained function

softmax function helps:

$$\mathcal{S}(\boldsymbol{a}) \triangleq \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}}, \ldots, \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}} \right]$$

$\mathcal{S}(\boldsymbol{a}/T)$ T is the temperature for the softmax



$\boldsymbol{a} = (3, 0, 1)$

Recap: Supervised Learning in a Nutshell

$\boldsymbol{x} \in \mathcal{X}$

inputs/features/covariates/predictors

model

f_θ

training objective/loss

$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$

training dataset

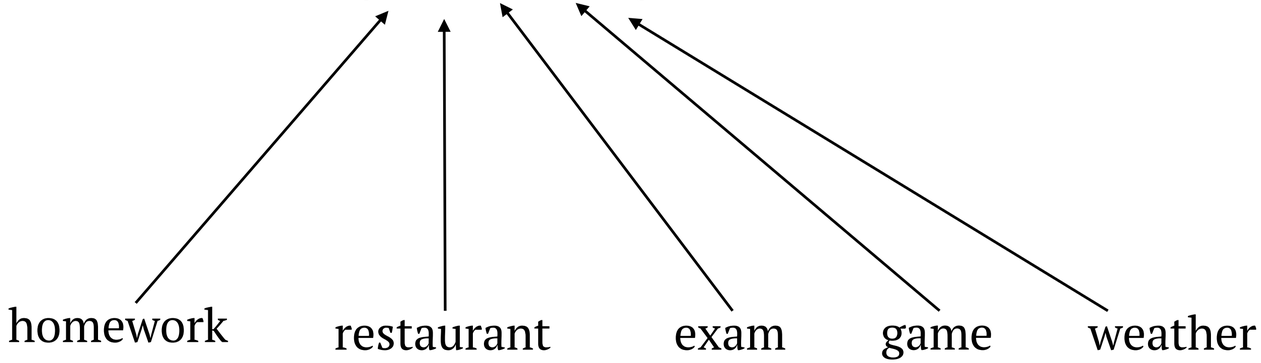
$\boldsymbol{y} \in \mathcal{Y}$

outputs/label/target/response

Recap: Supervised Learning in a Nutshell

Happy or sad? $y = 0, 1$

$$\boldsymbol{x} = \{0, 1, 0, 1, 0\}$$



$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \text{Ber}(y \mid \sigma(f(\boldsymbol{x}; \boldsymbol{\theta})))$$

sigmoid

$$D \quad \begin{array}{ll} \boldsymbol{x} = \{0, 0, 1, 1, 1\} & y = 1 \\ \boldsymbol{x} = \{0, 1, 1, 1, 1\} & y = 1 \\ \boldsymbol{x} = \{0, 1, 1, 0, 1\} & y = 0 \\ \vdots & \vdots \\ \boldsymbol{x} = \{1, 1, 1, 0, 1\} & y = 0 \end{array}$$

$$P(D \mid \boldsymbol{\theta}) = \prod_{i=0}^{|D|} p(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})$$

$$= \prod_{i=0}^{|D|} \text{Ber}(y_i \mid \sigma(f(\boldsymbol{x}_i; \boldsymbol{\theta})))$$

Recap: Supervised Learning in a Nutshell

$$P(D \mid \boldsymbol{\theta}) = \prod_{i=0}^{|D|} \text{Ber}(y_i \mid \sigma(f(\boldsymbol{x}_i; \boldsymbol{\theta})))$$

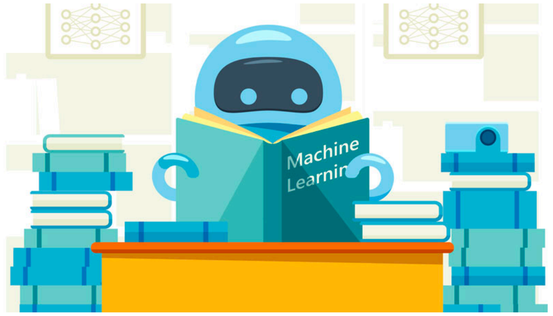
$$\boldsymbol{\sigma}(a) \triangleq \frac{1}{1+e^{-a}}$$

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + b$$

$$\boldsymbol{\theta} = \{w_1, w_2, w_3, w_4, w_5, b\}$$ parameters

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

The machine, is learning!



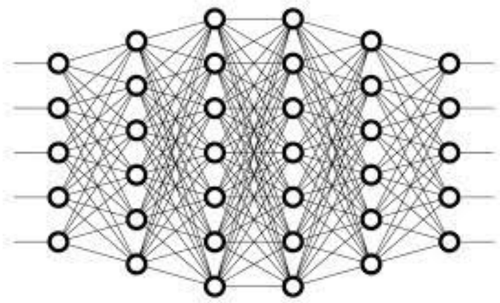
Recap: Supervised Learning in a Nutshell

$$P(D \mid \boldsymbol{\theta}) = \prod_{i=0}^{|D|} \text{Ber}(y_i \mid \sigma(f(\boldsymbol{x}_i; \boldsymbol{\theta})))$$

$$\boldsymbol{\sigma}(a) \triangleq \frac{1}{1+e^{-a}}$$

multinomial, HMMs…

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + b$$



$$\boldsymbol{\theta} = \{w_1, w_2, w_3, w_4, w_5, b\}$$ parameters

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

MLE, MAP …

Gradient Descent, SGD, Adam …

Recap: Supervised Learning in a Nutshell