

BBM406

Fundamentals of Machine Learning

Lecture 7:
Probability Review (cont'd.)
Maximum Likelihood Estimation (MLE)

Administrative

- **Project proposal** due November 15
- A half page description
 - problem to be investigated,
 - why it is interesting,
 - what data you will use,
 - related work.

A dramatic scene from a movie or video game. A Tyrannosaurus Rex is shown in mid-stride, running towards the viewer. The dinosaur is brown with dark stripes and has its mouth open, showing sharp teeth. It is running on a paved road that curves through a landscape with green hills and a cloudy sky. In the foreground, the side view mirror and door handle of a dark-colored car are visible, suggesting the viewer is inside the vehicle. The overall atmosphere is one of a high-stakes chase or confrontation.

Deadlines in the syllabus are
closer than they appear

Today

- Probabilities
 - Dependence, Independence, Conditional Independence
- Parameter estimation
 - Maximum Likelihood Estimation (MLE)
 - Maximum a Posteriori (MAP)

Last time... Sample space

Def: A **sample space** Ω is the set of all possible outcomes of a (conceptual or physical) random experiment. (Ω can be finite or infinite.)

Examples:

- Ω may be the set of all possible outcomes of a dice roll (1,2,3,4,5,6)
- Pages of a book opened randomly. (1-157)
- Real numbers for temperature, location, time, etc

Last time... Events

We will ask the question:

What is the probability of a particular event?

Def: Event A is a **subset** of the sample space Ω

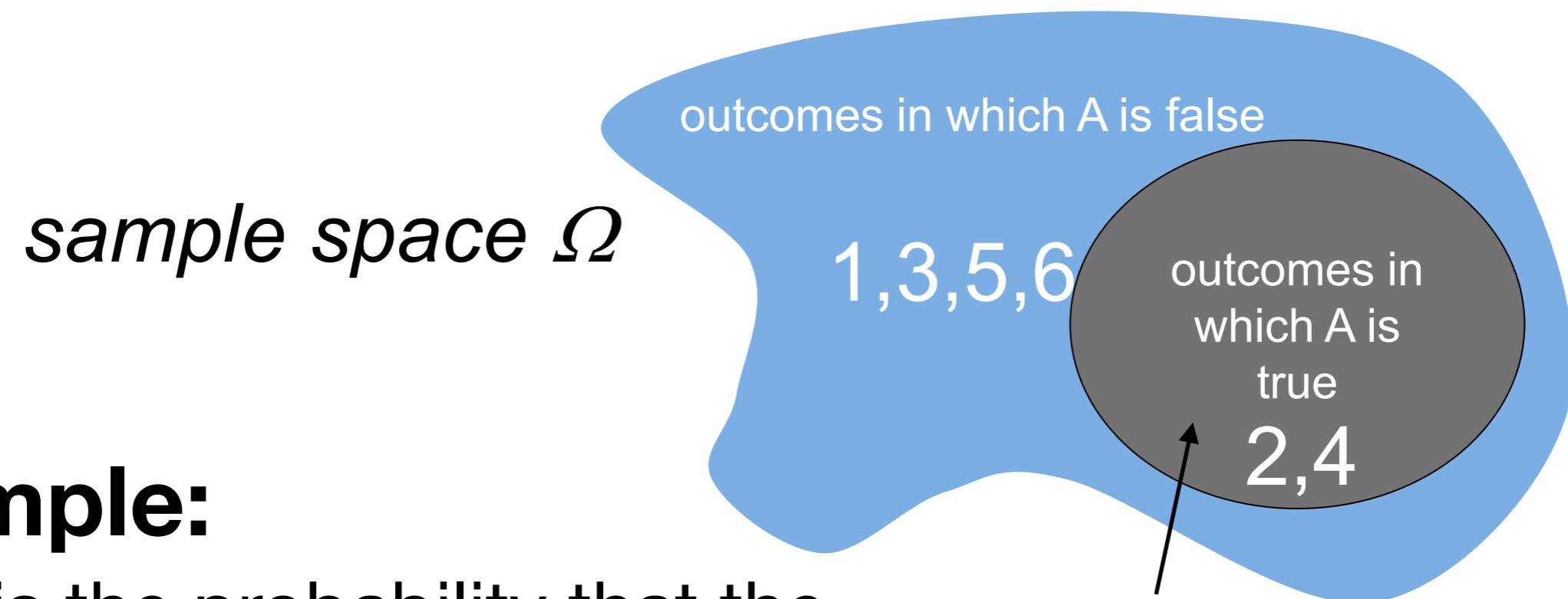
Examples:

What is the probability of

- the book is open at an odd number
- rolling a dice the number <4
- a random person's height X : $a < X < b$

Last time... Probability

Def: Probability $P(A)$, the probability that event (subset) A happens, is a function that maps the event A onto the interval $[0, 1]$. $P(A)$ is also called the **probability measure** of A .



Example:

What is the probability that the number on the dice is 2 or 4?

$P(A)$ is the volume of the area.

Last time... Kolmogorov Axioms

- (i) Nonnegativity: $P(A) \geq 0$ for each A event.
- (ii) $P(\Omega) = 1$.
- (iii) σ -additivity: For disjoint sets (events) A_i , we have

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

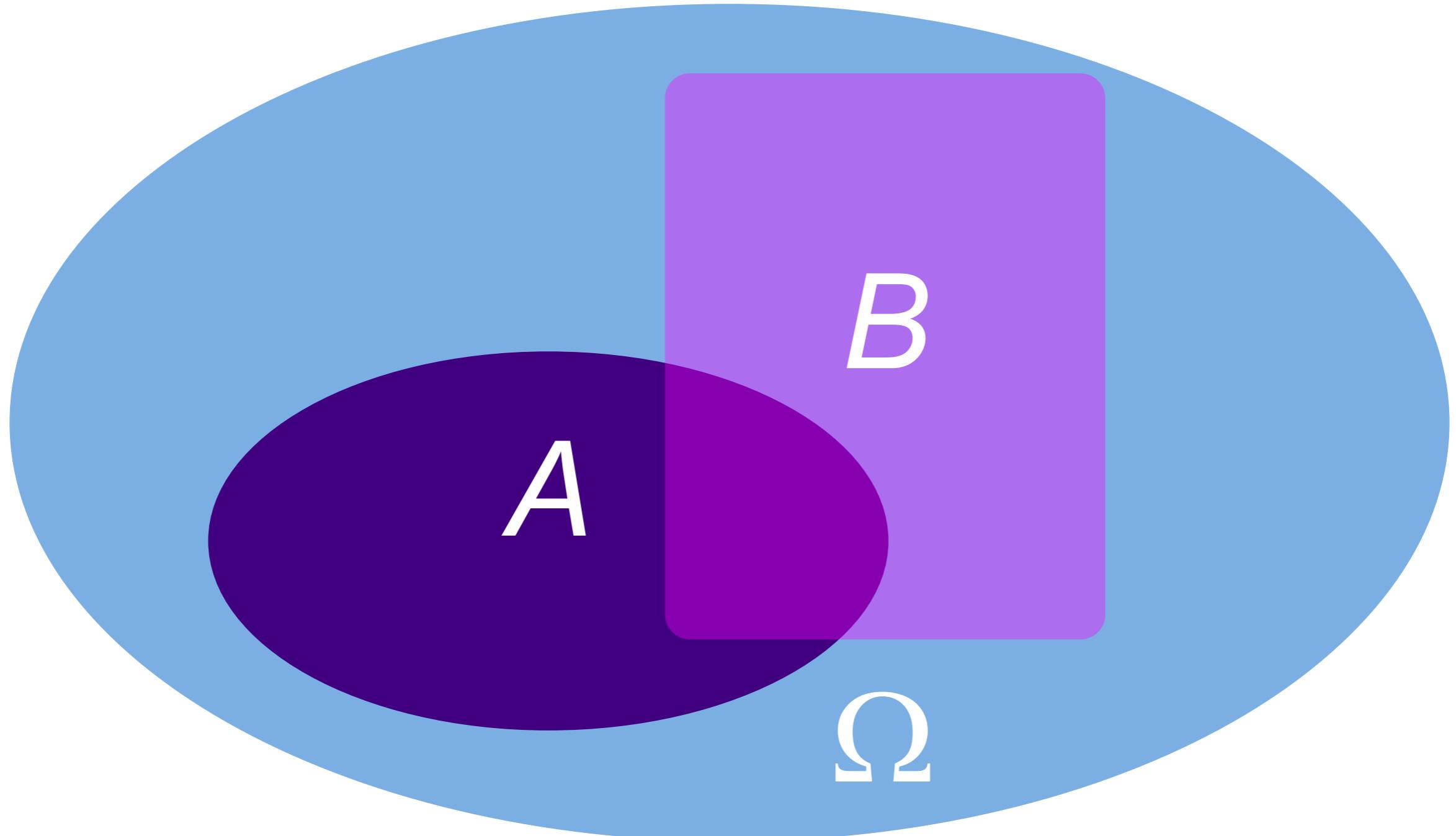
Consequences:

$$P(\emptyset) = 0.$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

$$P(A^c) = 1 - P(A).$$

Last time... Venn Diagram



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Last time... Random Variables

Def: Real valued **random variable** is a function of the outcome of a randomized experiment

$$X : \Omega \rightarrow \mathbb{R}$$

$$P(a < X < b) \doteq P(\omega : a < X(\omega) < b)$$

$$P(X = a) \doteq P(\omega : X(\omega) = a)$$

Examples:

- **Discrete random variable examples (Ω is discrete):**
- $X(\omega) = \text{True}$ if a randomly drawn person (ω) from our class (Ω) is female
- $X(\omega) = \text{The hometown } X(\omega) \text{ of a randomly drawn person } (\omega) \text{ from our class } (\Omega)$

Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

$\Omega = \{\text{head, tail}\}$ $X(\text{head}) = 1, X(\text{tail}) = 0.$

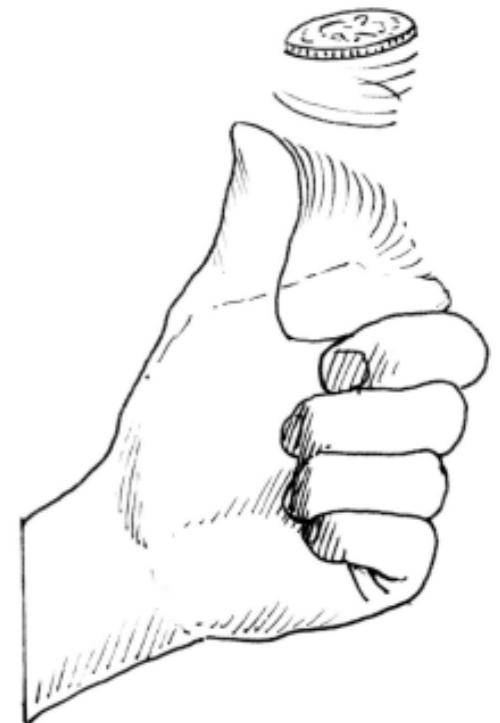


Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

$\Omega = \{\text{head, tail}\}$ $X(\text{head}) = 1, X(\text{tail}) = 0.$

$$P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases}$$



Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

$\Omega = \{\text{head, tail}\}$ $X(\text{head}) = 1, X(\text{tail}) = 0.$

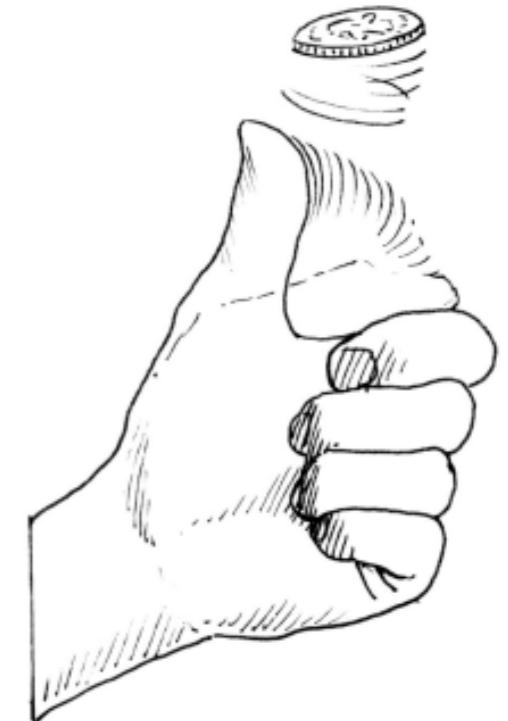
$$P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases}$$

- Binomial distribution: $\text{Bin}(n,p)$

Suppose a coin with head prob. p is tossed n times. What is the probability of getting k heads and $n-k$ tails?

$\Omega = \{ \text{possible } n \text{ long head/tail series}\}, |\Omega| = 2^n$

$K(\omega) = \text{number of heads in } \omega = (\omega_1, \dots, \omega_n) \in \{\text{head, tail}\}^n = \Omega$



Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

$\Omega = \{\text{head, tail}\}$ $X(\text{head}) = 1, X(\text{tail}) = 0.$

$$P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases}$$

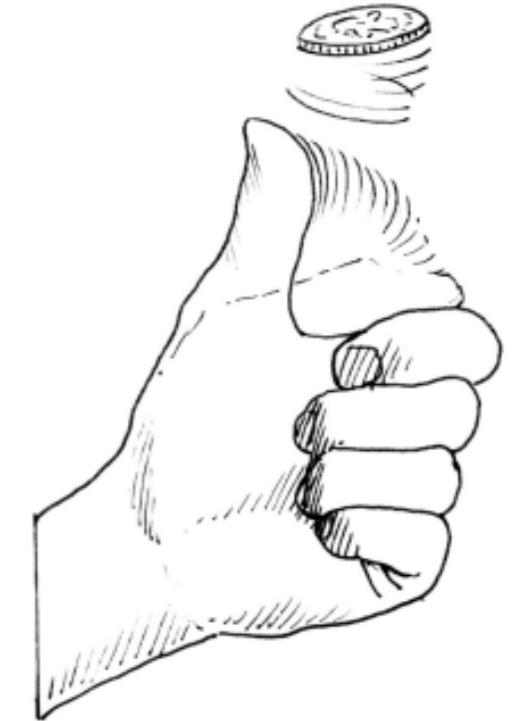
- Binomial distribution: $\text{Bin}(n,p)$

Suppose a coin with head prob. p is tossed n times. What is the probability of getting k heads and $n-k$ tails?

$\Omega = \{ \text{possible } n \text{ long head/tail series}\}, |\Omega| = 2^n$

$K(\omega) = \text{number of heads in } \omega = (\omega_1, \dots, \omega_n) \in \{\text{head, tail}\}^n = \Omega$

$$P(K = k) = P(\omega : K(\omega) = k) = \sum_{\omega : K(\omega) = k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$



Last time... Conditional Probability

$P(X|Y)$ = Fraction of worlds in which X event is true given Y event is true.

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

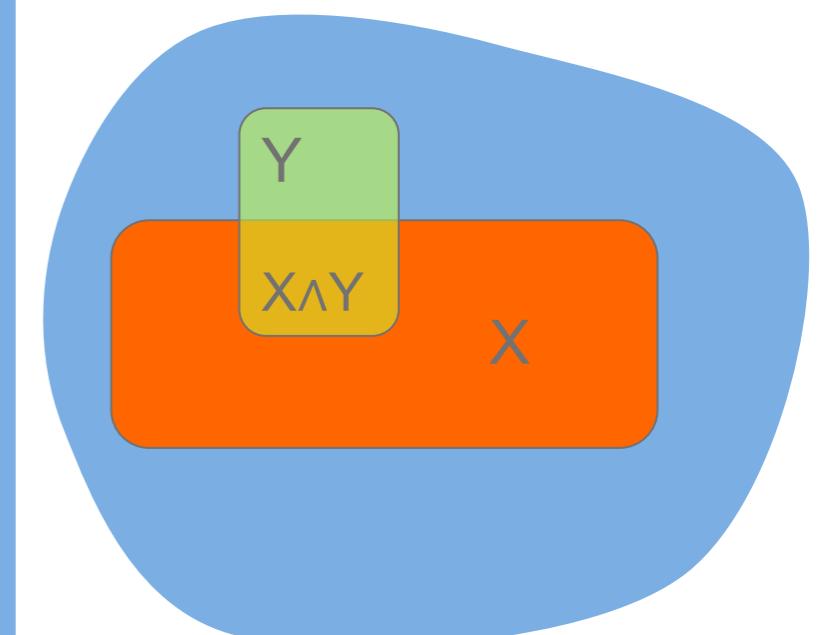
Last time... Conditional Probability

$P(X|Y)$ = Fraction of worlds in which X event is true given Y event is true.

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

$$P(\text{flu}|\text{headache}) = \frac{P(\text{flu, headache})}{P(\text{headache})} = \frac{1/80}{1/80 + 7/80}$$

	Flu	No Flu
Headache	1/80	7/80
No Headache	1/80	71/80



Independence

Independent random variables:

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

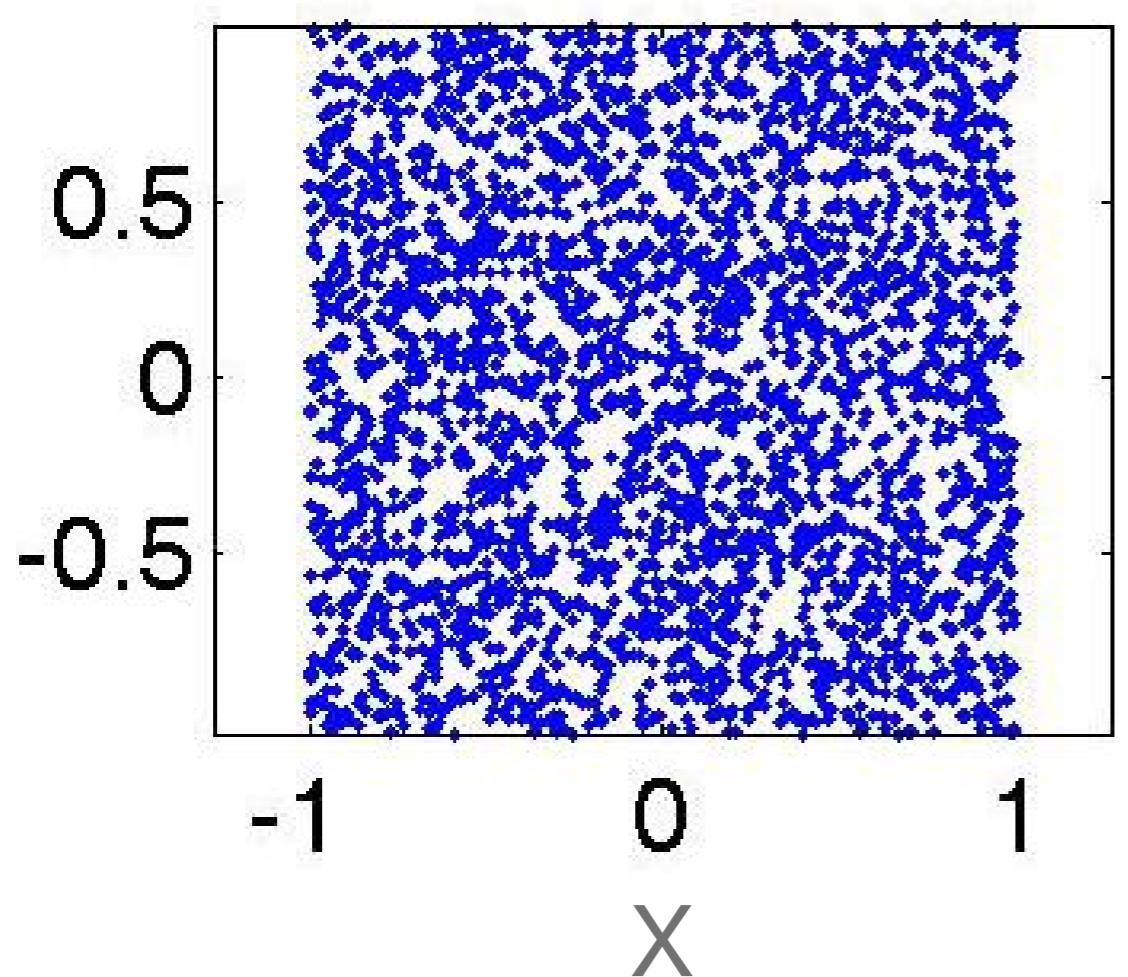
$$P(X|Y) = P(X)$$

Y and X don't contain information about each other.
Observing Y doesn't help predicting X.
Observing X doesn't help predicting Y.

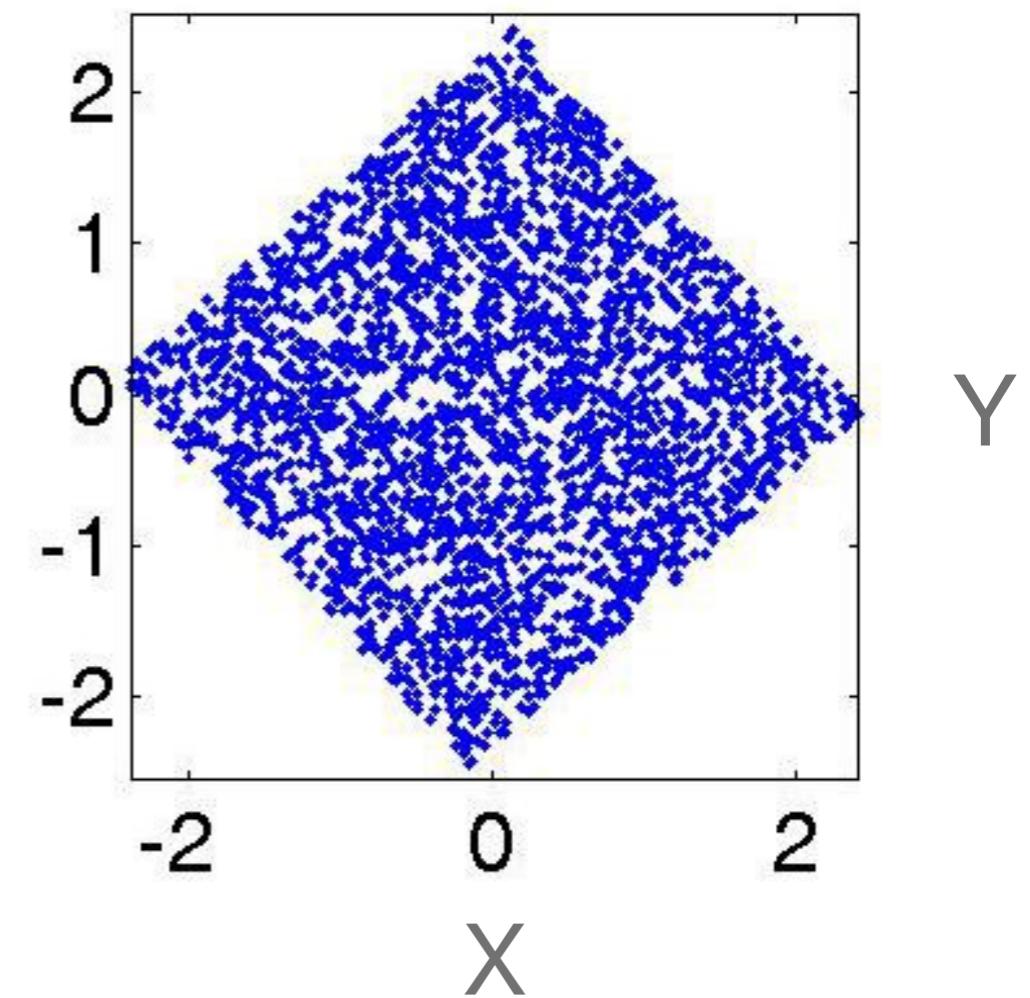
Examples:

Independent: Winning on roulette this week and next week.
Dependent: Russian roulette

Dependent / Independent



Independent X,Y



Dependent X,Y

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

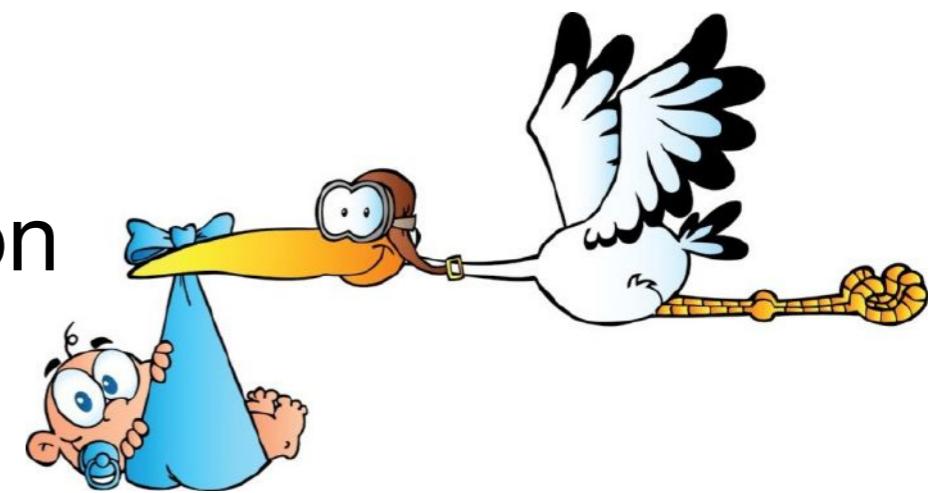
Examples:

Dependent: shoe size of children and reading skills

Conditionally independent: shoe size of children and reading skills given **age**

Stork deliver babies:

Highly statistically significant correlation exists between stork populations and human birth rates across Europe.



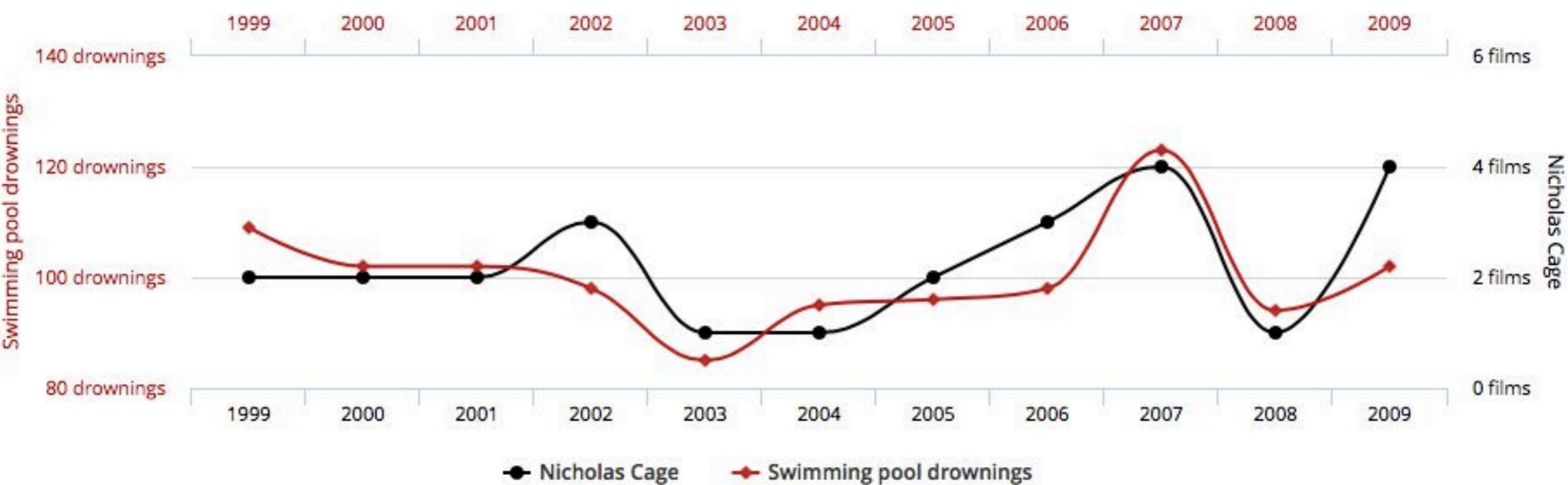
Conditionally Independent

- **London taxi drivers:** A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally, another study pointed out that people wear coats when it rains...

Correlation ≠ Causation

**Number people who drowned by falling into a swimming-pool correlates with
Number of films Nicolas Cage appeared in**



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Correlation: 0.666004

Conditional Independence

Formally: X is **conditionally independent** of Y given Z

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

$$P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain})$$

Conditional Independence

Formally: X is **conditionally independent** of Y given Z

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

$$P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain})$$

Equivalent to:

$$(\forall x, y, z)P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

Conditional Independence

Formally: X is **conditionally independent** of Y given Z

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

$$P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain})$$

Equivalent to:

$$(\forall x, y, z)P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

$$P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$$

Note: does NOT mean Thunder is independent of Rain

But given Lightning knowing Rain doesn't give more info about Thunder

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?

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:

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:



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, $D =$



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

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

MLE for Bernoulli distribution

Data, $D =$



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

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

Flips are i.i.d.:

MLE for Bernoulli distribution

Data, $D =$



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

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

Flips are i.i.d.:

- Independent events
 - Identically distributed according to Bernoulli distribution

MLE for Bernoulli distribution

Data, $D =$



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

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

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

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D|\theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i|\theta)\end{aligned}$$

independent draws

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D|\theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i|\theta) \quad \text{independent draws} \\ &= \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta) \quad \text{identically distributed}\end{aligned}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

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

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

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

$$\alpha_H (1-\theta) - \alpha_T \theta \Big|_{\theta=\hat{\theta}_{MLE}} = 0$$

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???**

Let θ^* be the true parameter.

For $n = a_H + a_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

For any $\epsilon > 0$:

Hoeffding's inequality:

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Probably Approximate Correct (PAC) Learning

I want to know the coin parameter θ , within $\epsilon = 0.1$ error with probability at least $1-\delta = 0.95$.

How many flips do I need?

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Sample complexity:

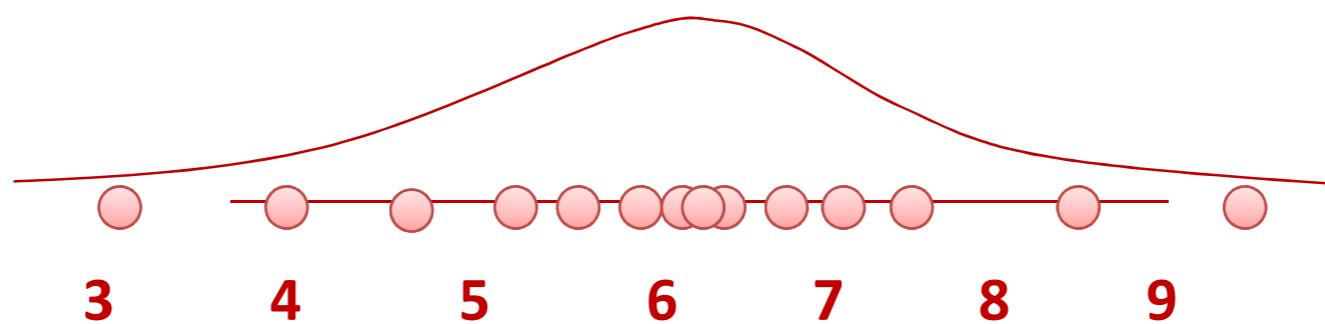
$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Question (3)

Why is this a machine learning problem???

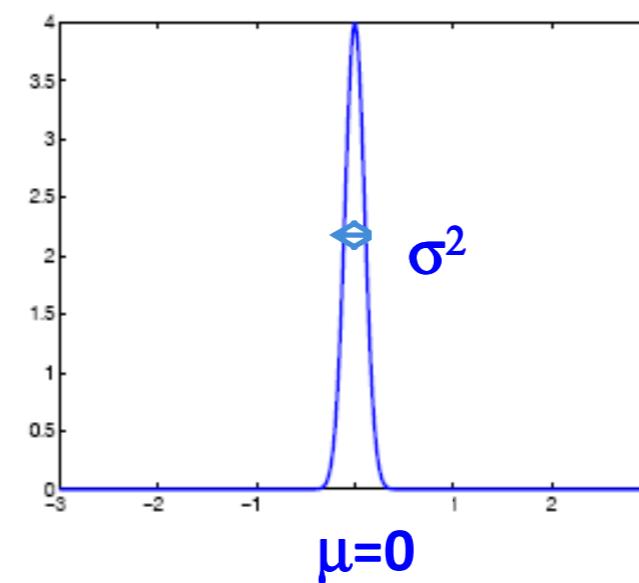
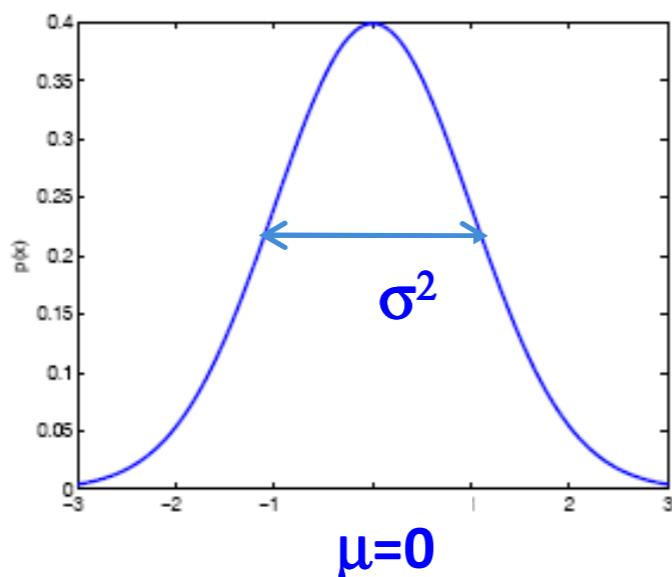
- improve their **performance** (accuracy of the predicted prob.)
- at some **task** (predicting the probability of heads)
- with **experience** (the more coins we flip the better we are)

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

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

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

Next Class:

MAP estimation
Naïve Bayes Classifier