CS182: Introduction to Machine Learning – MLE, MAP

Yujiao Shi
SIST, ShanghaiTech

Spring, 2025

Slides Courtesy of Matt Gormley & Henry Chai, CMU

Solving Linear Regression

上海科技大学
ShanghaiTech University

Question:

True or False: If Mean Squared Error (i.e. $\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - h(\mathbf{x}^{(i)}))^2$) has a unique minimizer (i.e. argmin), then Mean Absolute Error (i.e. $\frac{1}{N}\sum_{i=1}^{N}|y^{(i)} - h(\mathbf{x}^{(i)})|$) must also have a unique minimizer.

Answer:

Independence

Independence

Independent random variables:

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

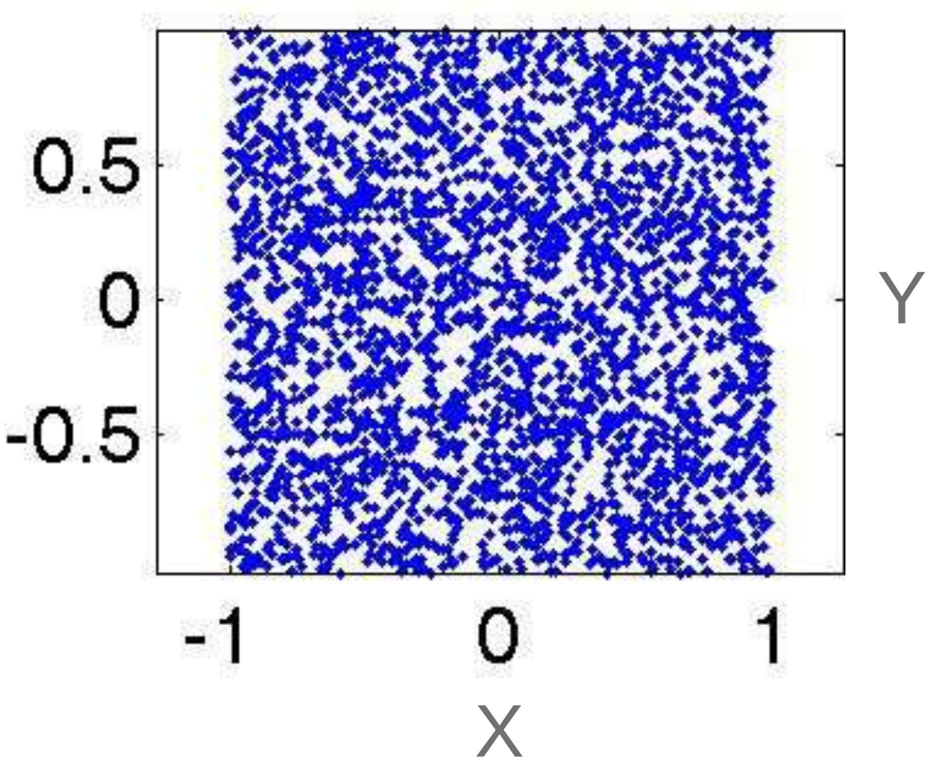
$$P(X|Y) = P(X)$$

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

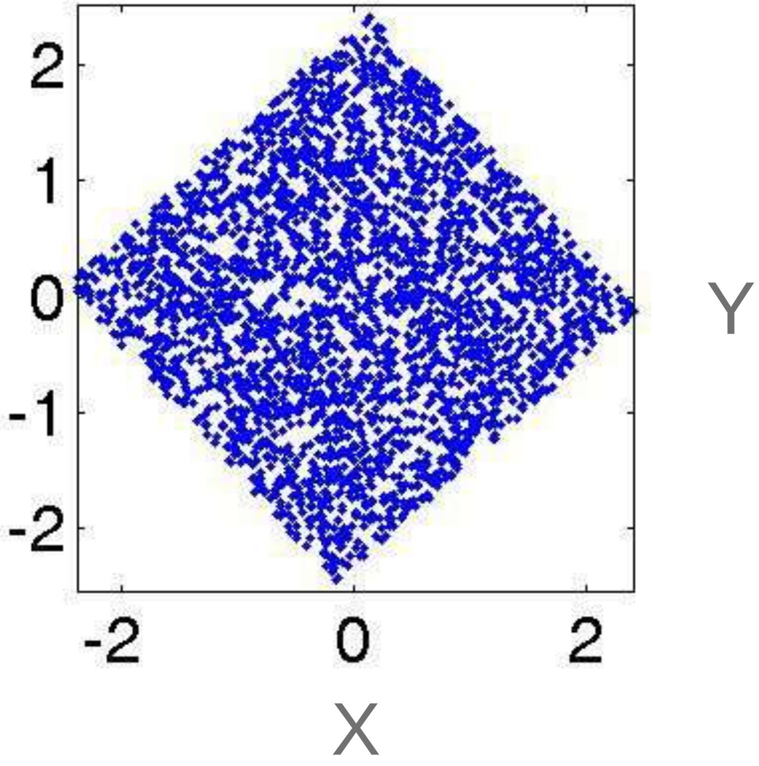
Examples:

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

Dependent / Independent



Independent X,Y



Dependent X,Y

Conditionally Independent

上海科技大学 ShanghaiTech University

Conditionally independent:

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

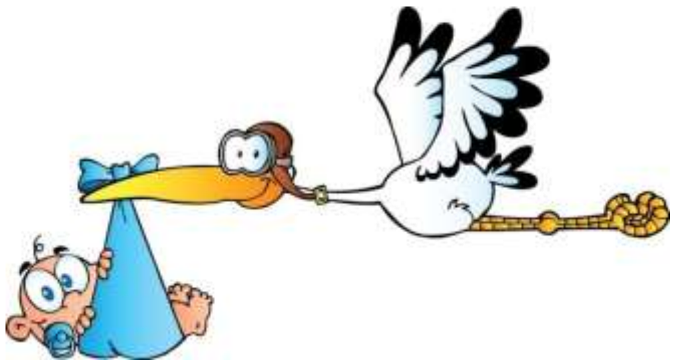
Knowing Z makes X and Y independent

Examples:

Dependent: shoe size and reading skills
Conditionally independent: shoe size and reading skills give n …?

Storks deliver babies:

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



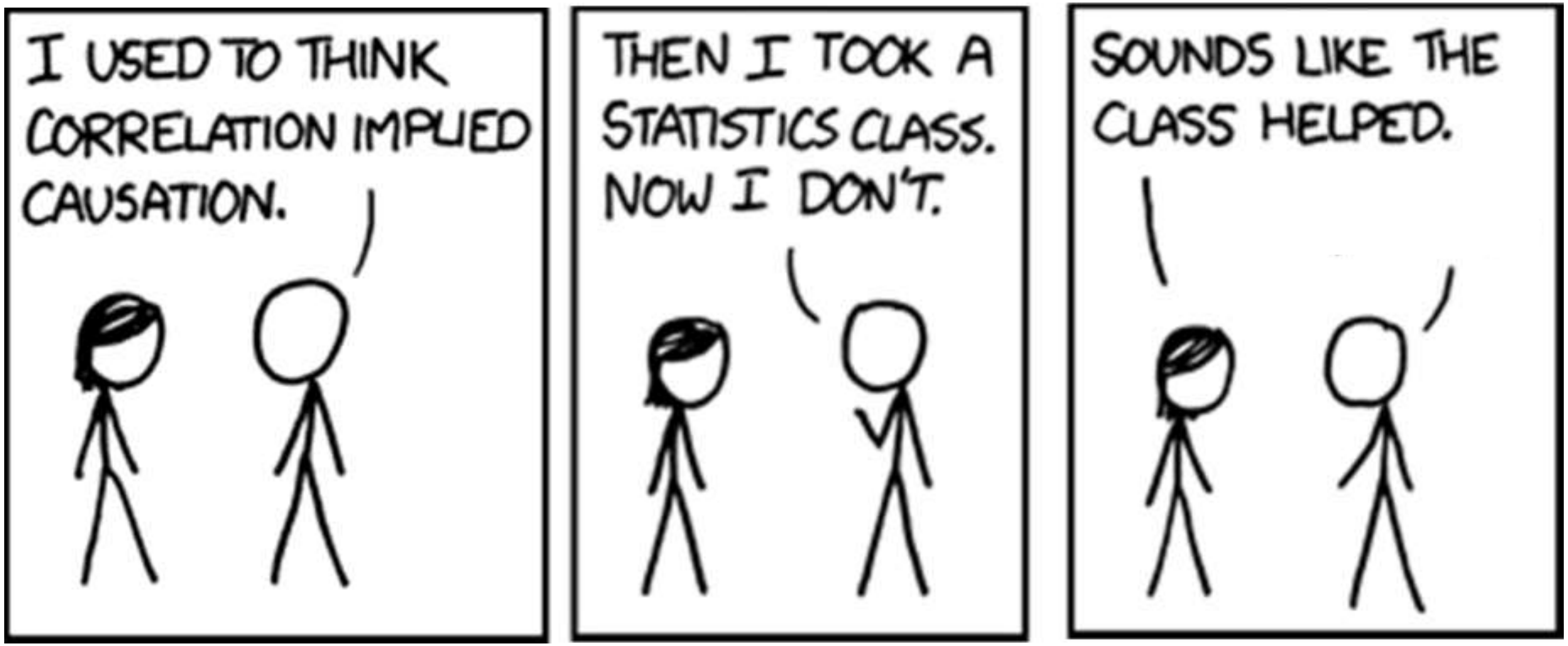
Conditionally Independent

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

上海科技大学
ShanghaiTech University



xkcd.com

Conditional Independence

上海科技大学 ShanghaiTech University

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

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

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

Equivalent to:

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

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

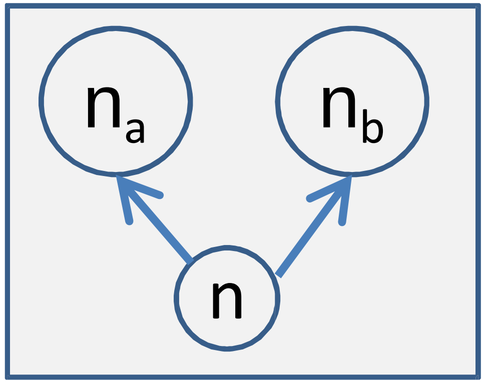
Note: does NOT mean Thunder is independent of Rain

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

Conditional vs. Marginal Independence

上海科技大学 ShanghaiTech University

- C calls A and B separately and tells them a number n ϵ {1,…,10}
- Due to noise in the phone, A and B each imperfectly (and independently) draw a conclusion about what the number was.
- A thinks the number was n_a and B thinks it was n_b.
- Are n_a and n_b marginally independent?
  - No, we expect e.g. $P(n_a = 1 \mid n_b = 1) > P(n_a = 1)$
- Are n_a and n_b conditionally independent given n?



  - Yes, because if we know the true number, the outcomes n_a and n_b are purely determined by the noise in each phone.

$$P(n_a = 1 \mid n_b = 1, n = 2) = P(n_a = 1 \mid n = 2)$$

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)

上海科技大学 ShanghaiTech University

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

Likelihood

- Given N independent, identically distribution (iid) samples $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$ of a random variable X
  - If X is discrete with probability mass function (pmf) $p(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$L(\theta) = \prod_{n=1}^{N} p\left(x^{(n)}|\theta\right)$$

  - If X is continuous with probability density function (pdf) $f(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$L(\theta) = \prod_{n=1}^{N} f\left(x^{(n)}|\theta\right)$$

MLE for Bernoulli distribution

Data, D =

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

$P(Heads) = \theta, \;\; P(Tails) = 1-\theta$

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

上海科技大学 ShanghaiTech University

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

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

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

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

That's exactly the "Frequency of heads"

Question (2)

上海科技大学 ShanghaiTech University

How good is this MLE estimation???

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

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???**

Simple Bound

Let θ^* be the true parameter.

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

For any $\varepsilon > 0$:

Hoeffding's inequality:

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

Probably Approximate Correct (PAC)Learning

上海科技大学
ShanghaiTech University

I want to know the coin parameter θ, within ε = 0.1 error with probability at least 1-δ = 0.95.

How many flips do I need?

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

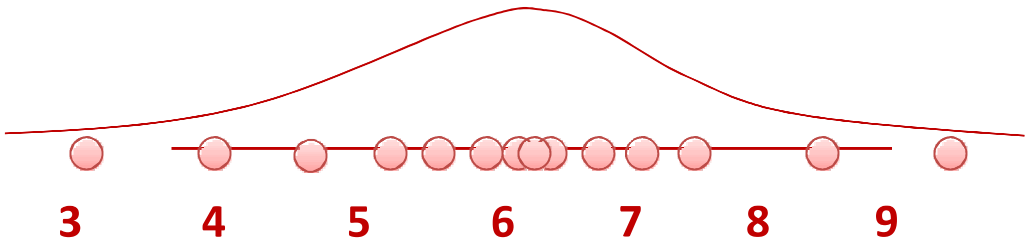
Sample complexity:

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

Question (3)

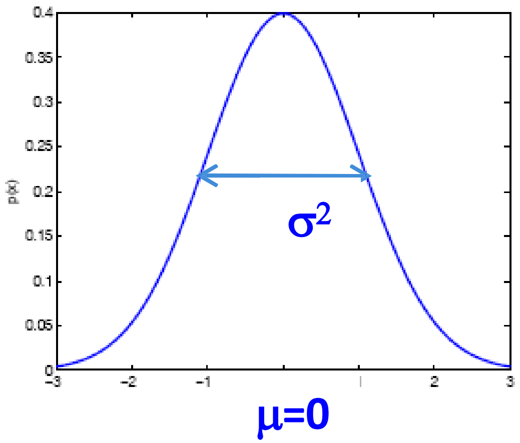
- Why is this a machine learning problem???
  - improve their performance (accuracy of the predicted prob.)
  - at some task (predicting the probability of heads)
  - with experience (the more coins we flip the better we are)

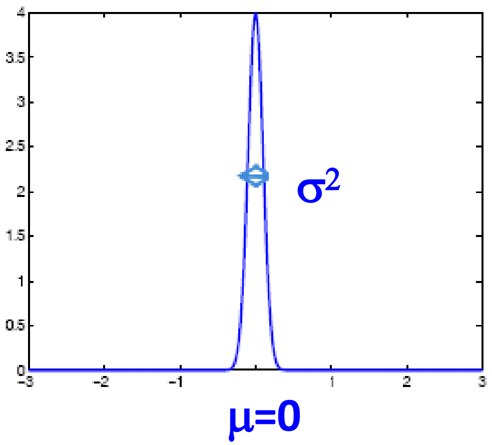
What about continuous features?



Let us try Gaussians...

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





MLE for Gaussian mean and variance

上海科技大学 ShanghaiTech University

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

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

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

MLE for Gaussian mean and variance

上海科技大学 ShanghaiTech University

$$
\begin{aligned}
\widehat{\mu}_{MLE} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\widehat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2
\end{aligned}
$$

上海科技大学
ShanghaiTech University

$$J(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

$$= \underbrace{(2\pi\sigma^2)^{-\frac{1}{2}n}}_{u}\underbrace{e^{-\alpha/2\sigma^2}}_{v}$$

$$\frac{\partial J}{\partial \mu} = J * \sum_{i=1}^{n}(X_i-\mu)/\sigma^2 = 0$$

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

$$\frac{\partial J}{\partial \sigma^2} = \frac{\partial u}{\partial \sigma^2} v + u\frac{\partial v}{\partial \sigma^2}$$

$$\frac{\partial u}{\partial \sigma^2} = -\frac{1}{2}n(2\pi\sigma^2)^{-\frac{1}{2}n-1} * 2\pi = -\frac{1}{2}n * u * \frac{1}{\sigma^2}$$

$$\frac{\partial v}{\partial \sigma^2} = v * \frac{\alpha}{2} * \frac{1}{\sigma^4}$$

$$\frac{\partial J}{\partial \sigma^2} = \frac{\partial u}{\partial \sigma^2} v + u\frac{\partial v}{\partial \sigma^2}$$

$$= -\frac{1}{2}n * u * \frac{1}{\sigma^2} * v + u * v * \frac{\alpha}{2} * \frac{1}{\sigma^4} = 0$$

$$-n + \alpha * \frac{1}{\sigma^2} = 0$$

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

MLE for Gaussian mean and variance

上海科技大学 ShanghaiTech University

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

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

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

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

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

Why MLE for the variance of a Gaussian is **biased?**

上海科技大学 ShanghaiTech University

Because it uses the sample mean $\hat{\mu}_{MLE}$ instead of the true mean μ

Consider the expectation of $\hat{\sigma}^2_{MLE}$:

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2\right]$$

It can be shown that:

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = \frac{n-1}{n}\sigma^2$$

Thus, the expectation of $\hat{\sigma}^2_{MLE}$ is smaller than the true variance σ^2, with the bias given by:

$$\mathrm{Bias} = \mathbb{E}[\hat{\sigma}^2_{MLE}] - \sigma^2 = -\frac{\sigma^2}{n}$$

上海科技大学
ShanghaiTech University

What about prior knowledge？

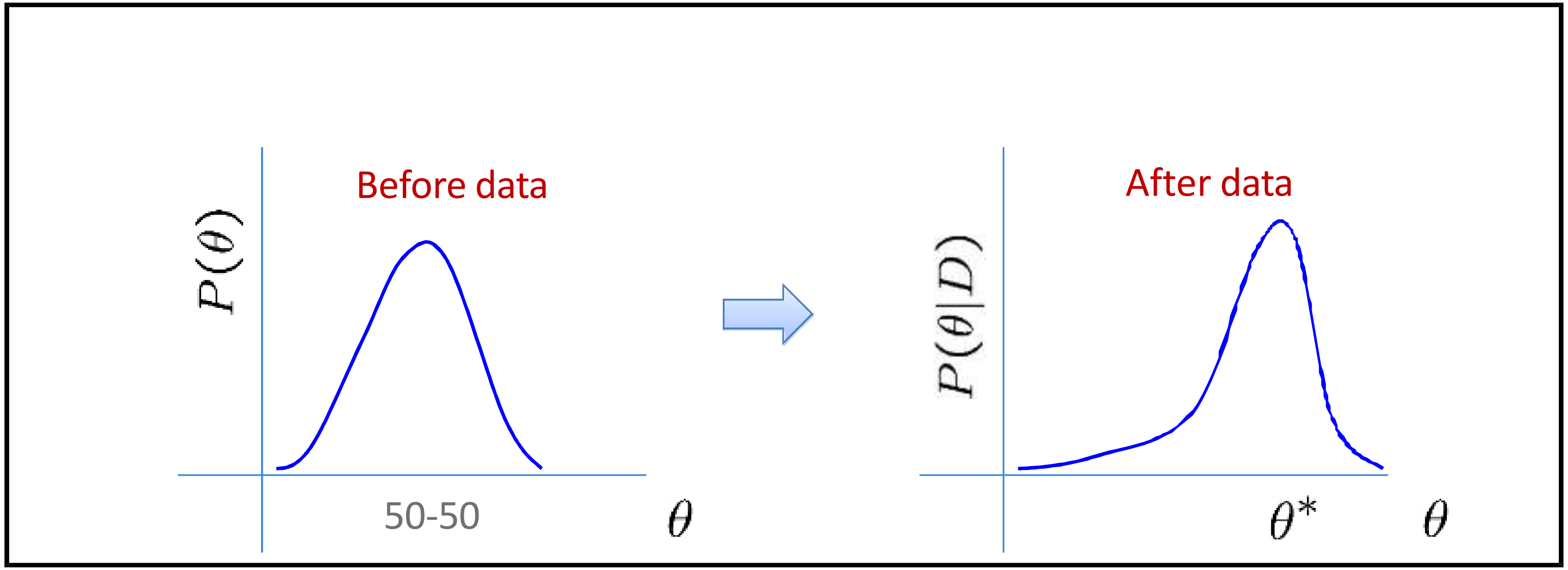
(MAP Estimation)

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



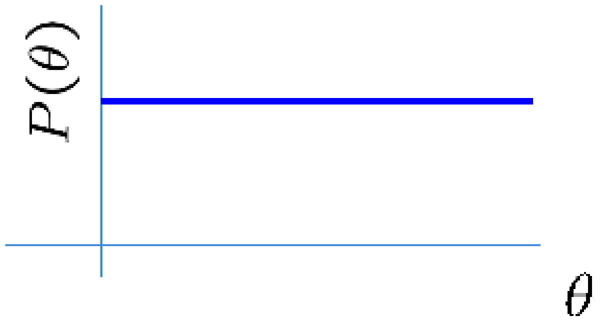
Prior distribution

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 $P(\theta|D)$ 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

上海科技大学
ShanghaiTech University

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

上海科技大学 ShanghaiTech University

- Use Bayes rule:

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

- Or equivalently:

$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

Maximum a Posteriori (MAP) Estimation

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters
  - MLE finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p\ (\mathcal{D}|\theta)$
  - MAP finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p\ (\theta|\mathcal{D})$

$$= \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}}\underbrace{p(\theta)}_{\text{prior}}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{\log p(\mathcal{D}|\theta) + \log p(\theta)}_{\text{log-posterior}}$$

ShanghaiTech University 上海科技大学

Okay, but how on earth do we pick a prior?

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters
  - MLE finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p\ (\mathcal{D}|\theta)$
  - MAP finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p\ (\theta|\mathcal{D})$

$$= \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}}\underbrace{p(\theta)}_{\text{prior}}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{\log p(\mathcal{D}|\theta) + \log p(\theta)}_{\text{log-posterior}}$$

MAP estimation for Binomial distribution

上海科技大学 ShanghaiTech University

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,

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

$\Rightarrow$ posterior is Beta distribution

Beta function: $B(\beta_H, \beta_T) = \int_0^1 \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} d\theta$
is a normalizing constant to ensure the distribution integrates to 1.

MAP estimation for Binomial distribution

The Beta distribution is the *conjugate prior* for the Bernoulli distribution!

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

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

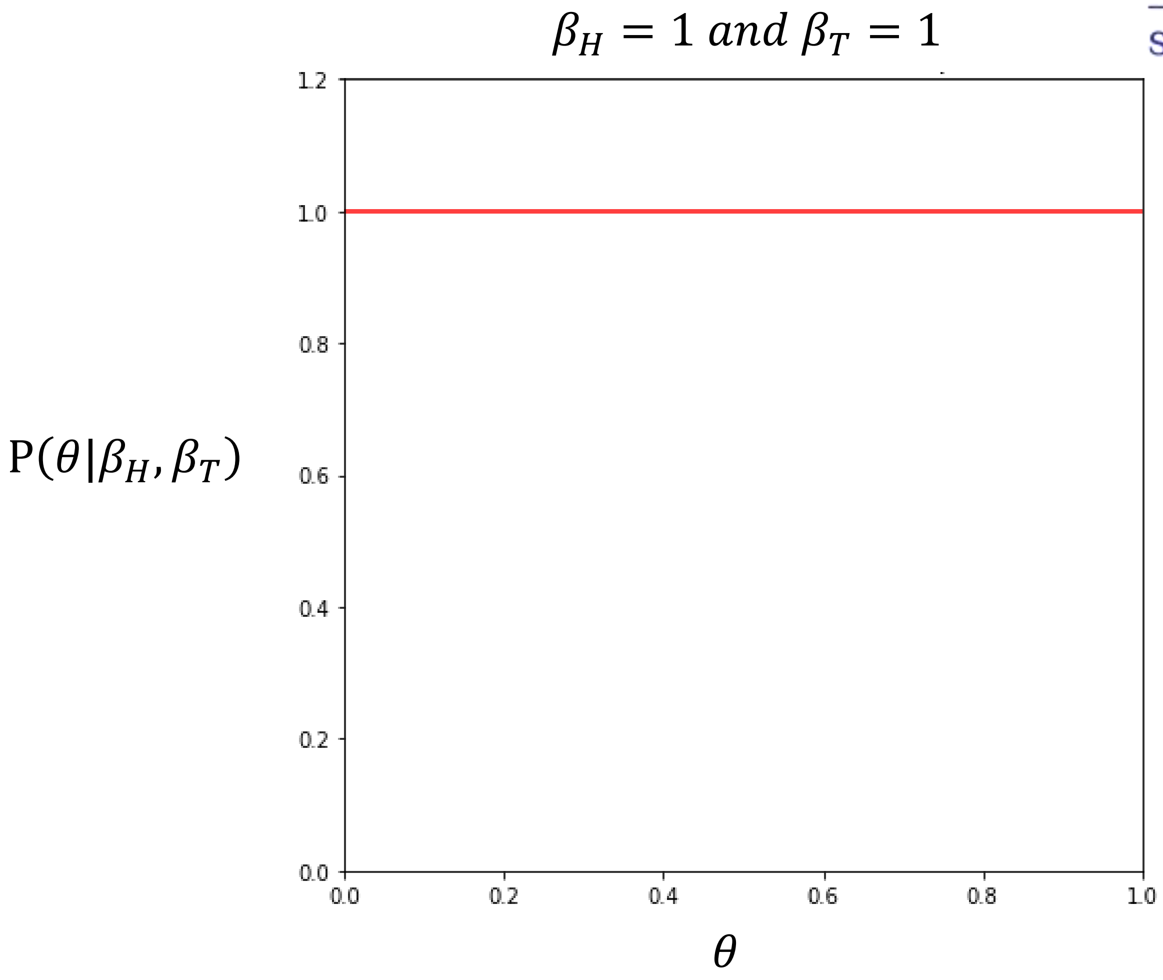
⇒ posterior is Beta distribution

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

P(θ) and P(θ|D) have the same form! [Conjugate prior]

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

Beta Distribution



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

MAP estimation for Binomial distribution

The Beta distribution is the *conjugate prior* for the Bernoulli distribution!

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

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

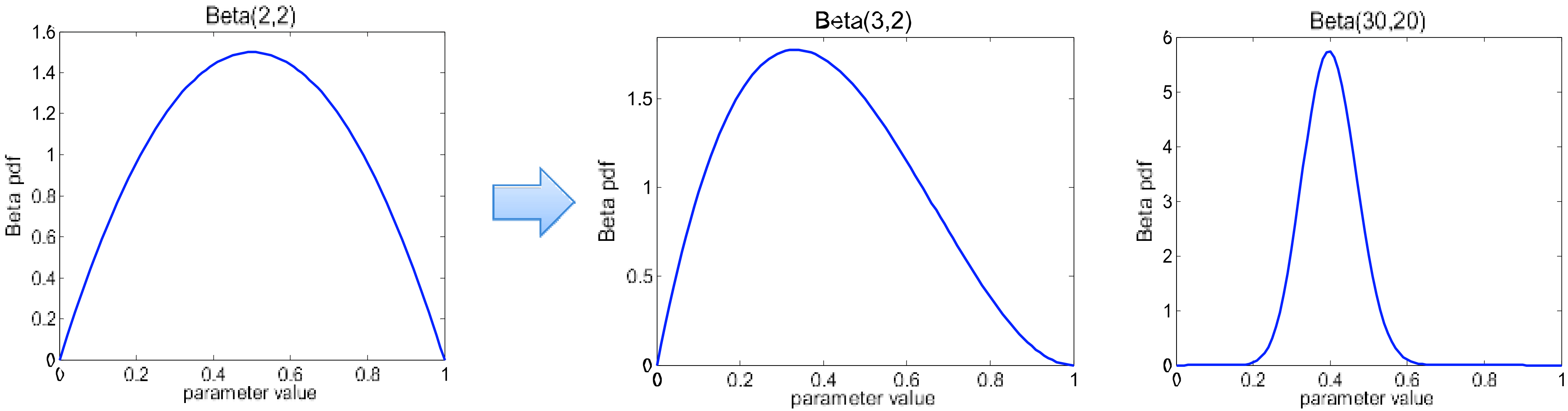
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ|D) have the same form! [Conjugate prior]

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

Beta conjugate prior

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



As $n = \alpha_H + \alpha_T$ increases

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

From Binomial to Multinomial

Example: 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,

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

Then posterior is Dirichlet distribution

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

For Multinomial, conjugate prior is Dirichlet distribution.

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

Bayesians vs.Frequentists

You are no good when sample is small

You give a different answer for different priors

MLE/MAP Learning Objectives

You should be able to…

- Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence
- State the principle of maximum likelihood estimation and explain what it tries to accomplish
- State the principle of maximum a posteriori estimation and explain why we use it
- Derive the MLE or MAP parameters of a simple model in closed form