Introduction to Machine Learning CMU-10701

2. MLE, MAP, Bayes classification

Barnabás Póczos & Aarti Singh

2014 Spring

Administration

http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/index.html

- Blackboard manager & Peer grading: Dani
- Webpage manager and autolab: Pulkit
- Camera man: Pengtao
- Homework manager: Jit
- Piazza manager: Prashant

Recitation: Wean 7500, 6pm-7pm, on Wednesdays

Outline

Theory:

- Probabilities:
  - Dependence, Independence, Conditional Independence
- Parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum aposteriori (MAP)
- Bayes rule
  - Naïve Bayes Classifier

Application:

Naive Bayes Classifier for

- Spam filtering
- "Mind reading" = fMRI data processing

Independence

Independence

Independent random variables:

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

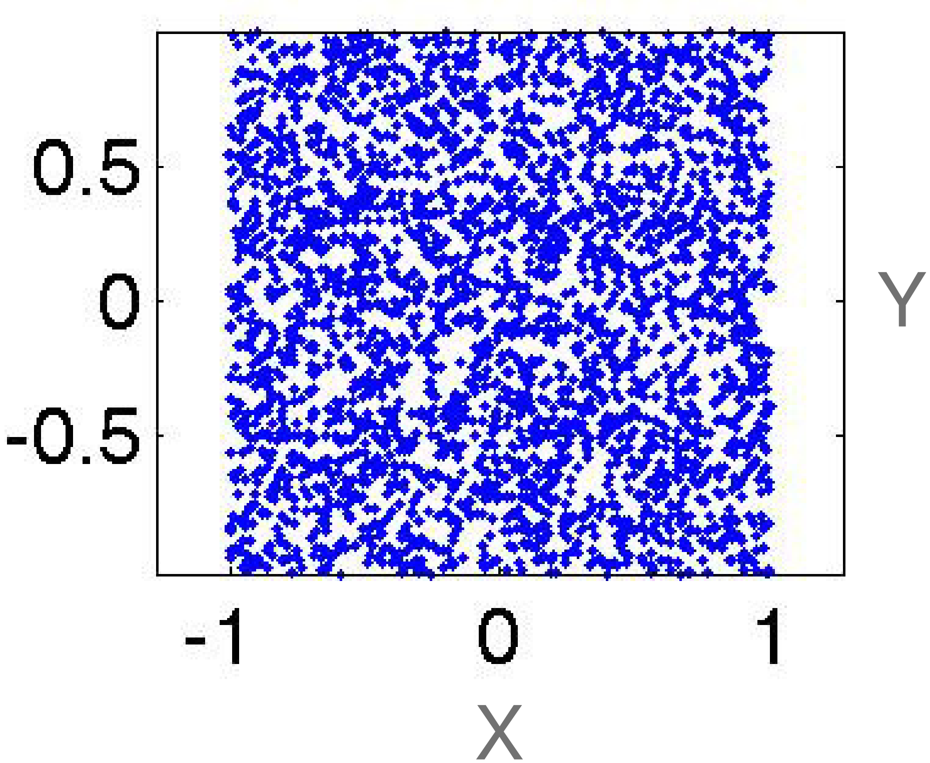
$$P(X|Y) = P(X)$$

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

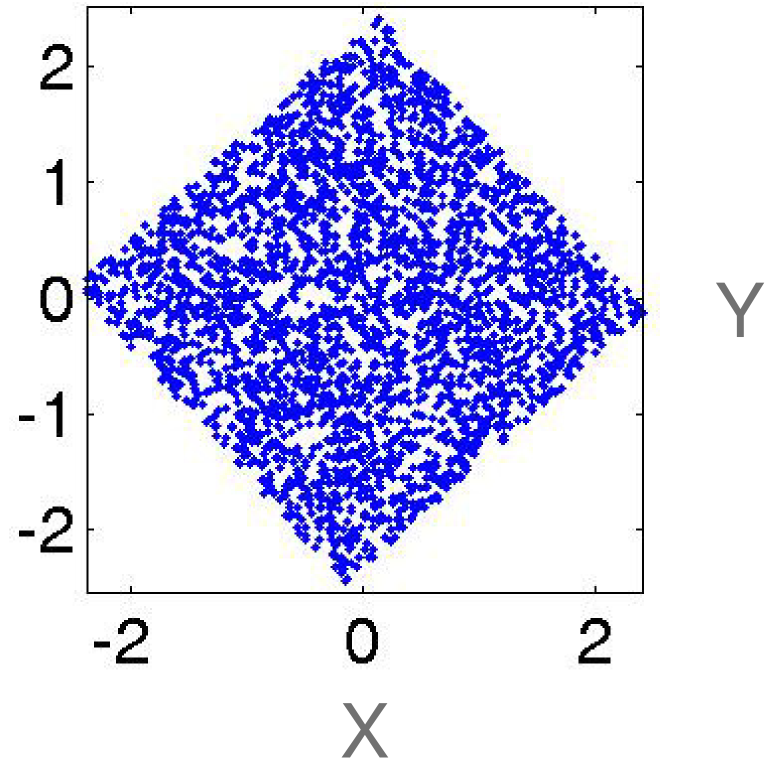
Examples:

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

Dependent / Independent



Independent X,Y



Dependent X,Y

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

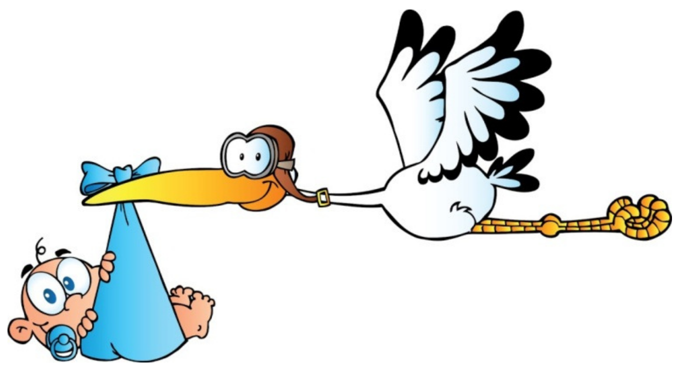
Examples:

Dependent: show size and reading skills

Conditionally independent: show size and reading skills given age

Storks deliver babies:

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

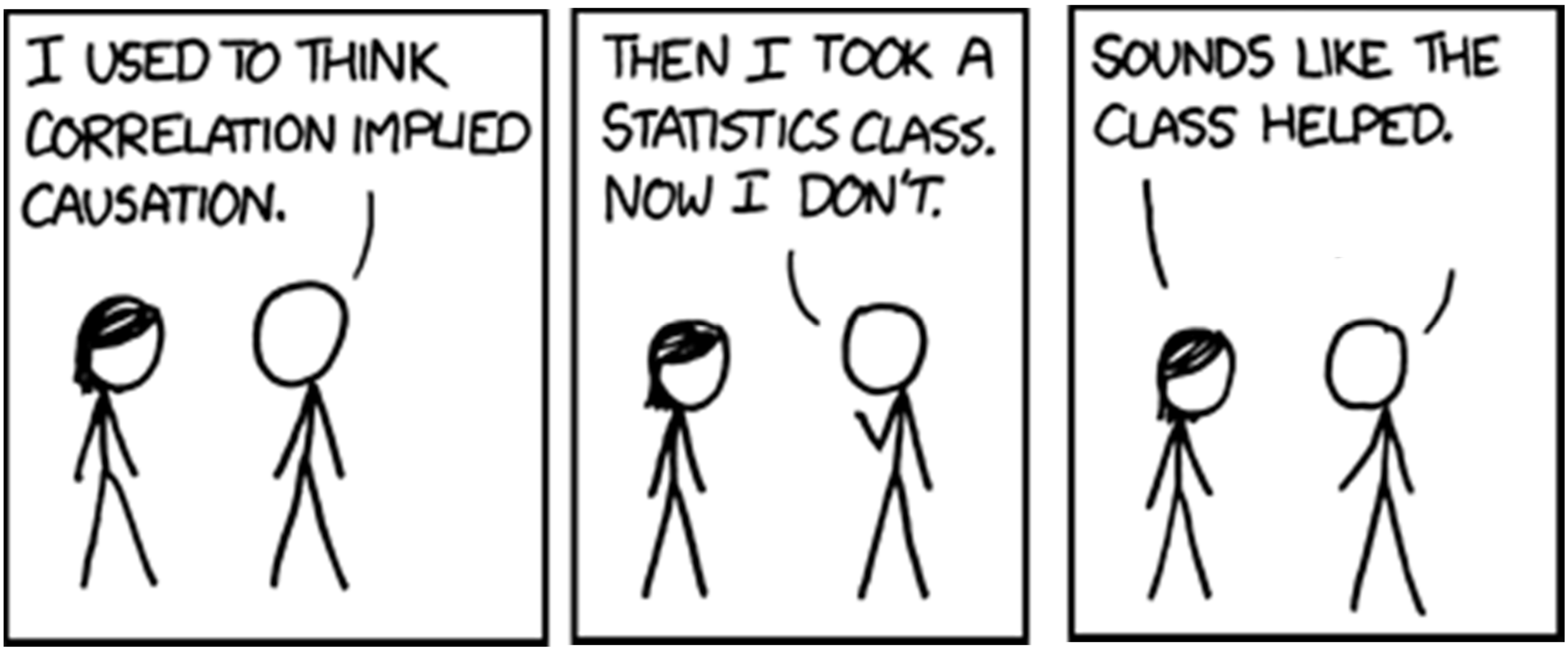


Conditionally Independent

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



xkcd.com

Conditional Independence

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

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

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

Equivalent to:

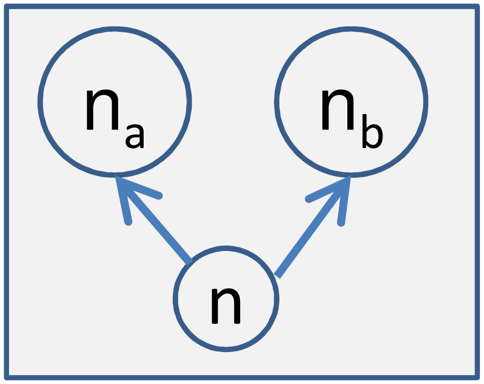
$$(\forall x, y, z)P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

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

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

- C calls A and B separately and tells them a number $n \in \{1,\ldots,10\}$
- 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.

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

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 =

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

P(Heads) = θ, *P(Tails)* = 1-θ

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

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta) \leftarrow P(x_1, x_2, \ldots x_n \mid \theta)$$

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

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

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

n
α_H
α_T

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

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

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

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

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

That's exactly the "Frequency of heads"

Question (2)

How good is this MLE estimation???

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

$\mathbb{E}\hat{\theta} = \theta$

$\text{Bias} = |\theta - \mathbb{E}\hat{\theta}|$

How many flips do I need?

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

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

What if I flipped 30 heads and 20 tails?

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

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

For any $\varepsilon > 0$:

Hoeffding's inequality:

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

Probably Approximate Correct (PAC)Learning

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?

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

0.05

$$e^{-2n\varepsilon^2} \leq \frac{\delta}{2}$$

$$-2n\varepsilon^2 \leq \ln\left(\frac{\delta}{2}\right)$$

$$\ln\left(\frac{2}{\delta}\right) \leq 2n\varepsilon^2 \implies n \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

Sample complexity:

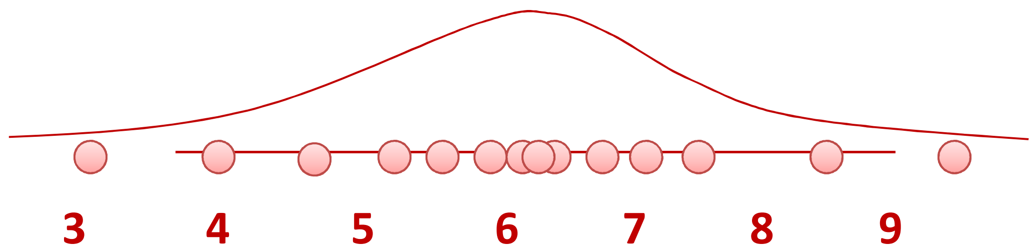
$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Question (3)

Why is this a machine learning problem???

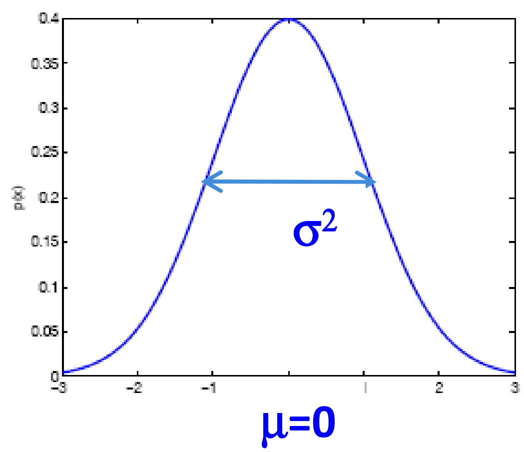
- improve their performance (accuracy of the predicted prob.)
- at some task (predicting the probability of heads)
- with experience (the more coins we flip the better we are)

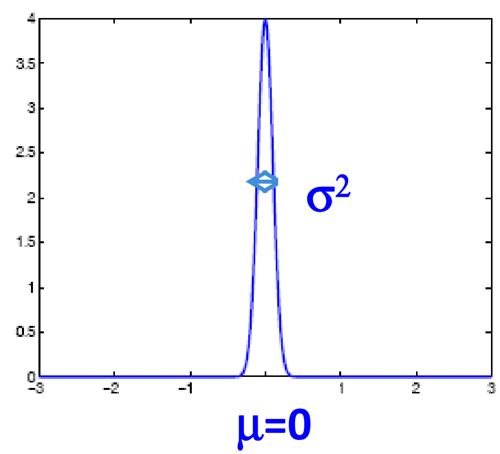
What about continuous features?



Let us try Gaussians…

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





MLE for Gaussian mean and variance

Choose θ= (μ,σ²) 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

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

Identically distributed

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

MLE for Gaussian mean and variance

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

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

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

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

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

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

What about prior knowledge?

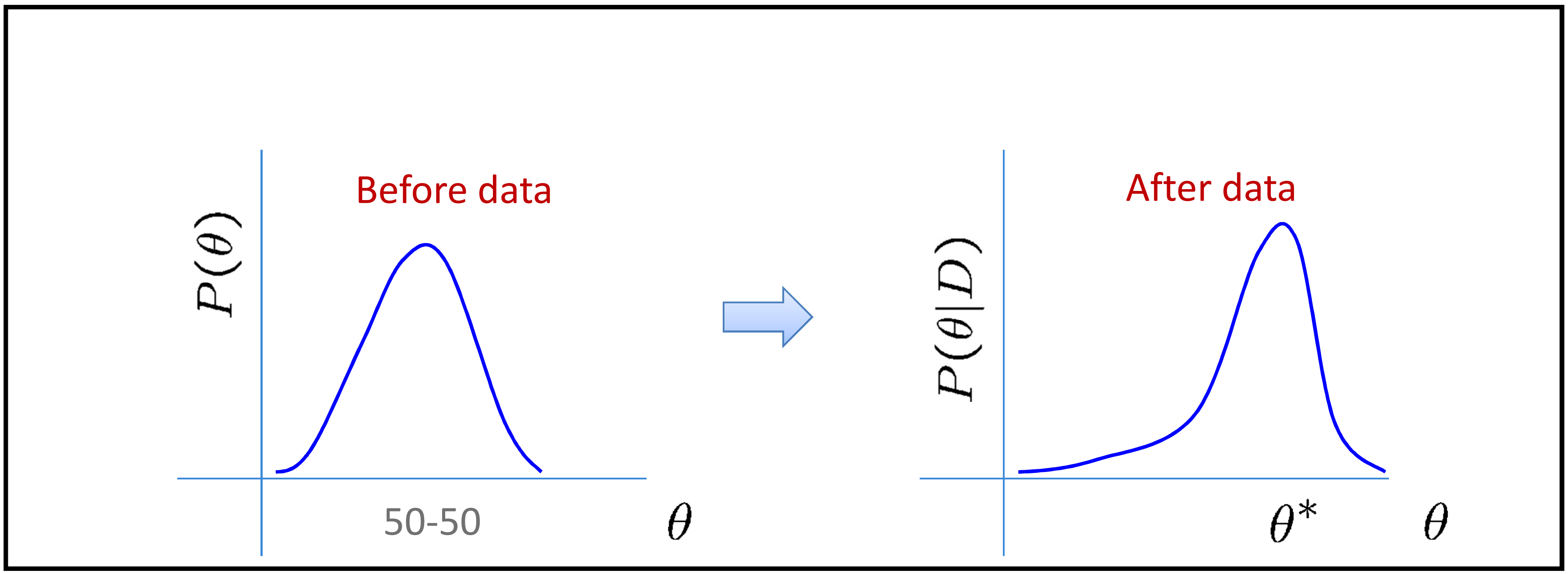
(MAP Estimation)

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



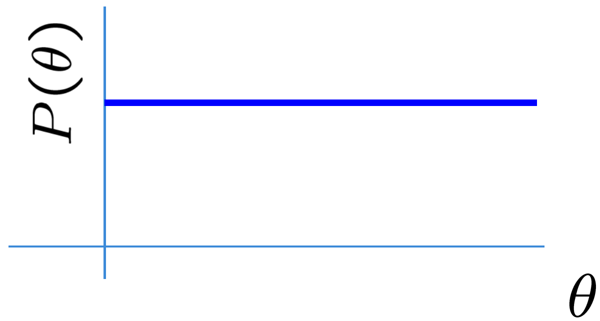
Prior distribution

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(θ) and P(θ|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

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:

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

- Or equivalently:

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

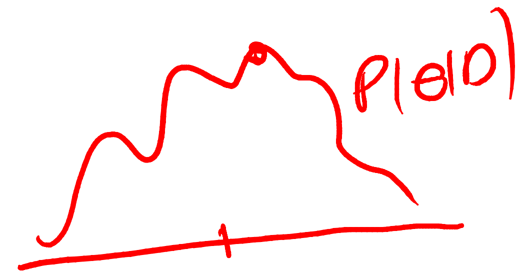
posterior likelihood prior

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

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



- Maximum *a posteriori* (MAP) estimation

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

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(D|\theta)P(\theta)$$

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,

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

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

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

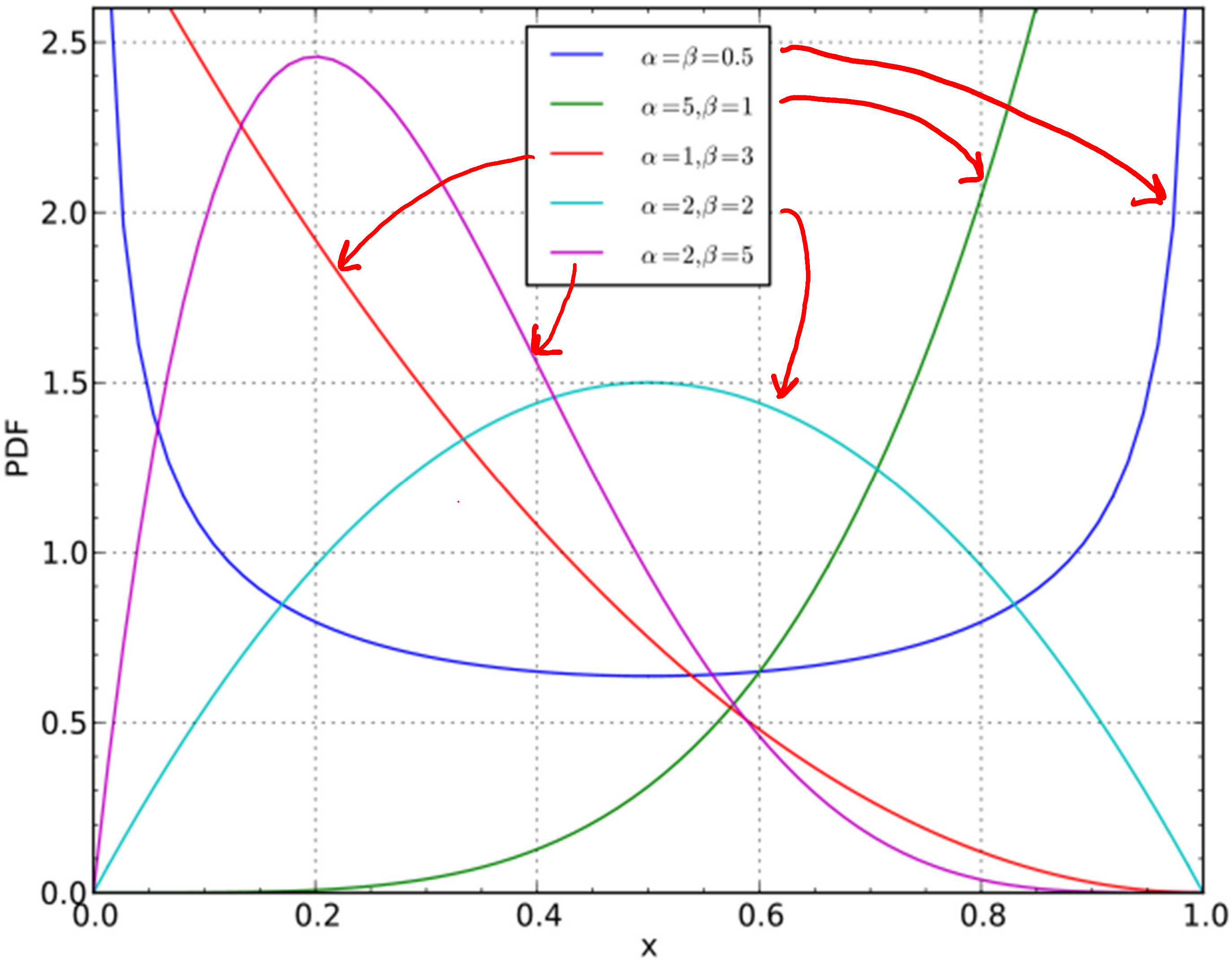
$\Rightarrow$ posterior is Beta distribution

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

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

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

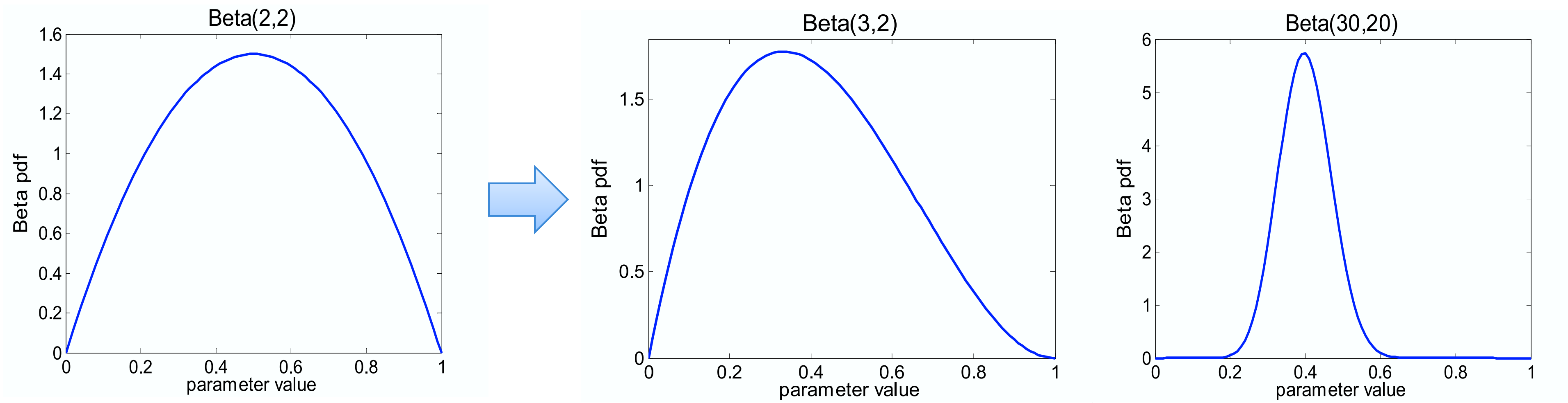
Beta distribution



More concentrated as values of α, β increase

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,

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

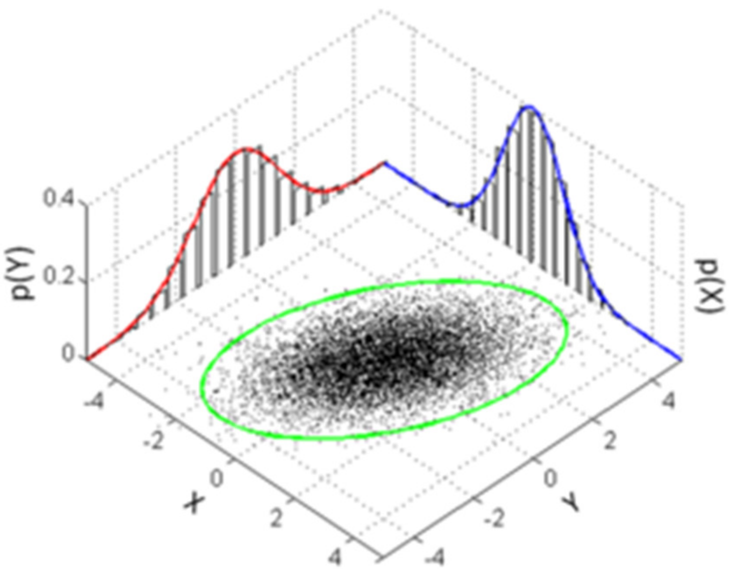
Then posterior is Dirichlet distribution

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

For Multinomial, conjugate prior is Dirichlet distribution.

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

Conjugate prior for Gaussian?



$$(2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

Conjugate prior on mean: **Gaussian**

Conjugate prior on covariance matrix: **Inverse Wishart**

$$\frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)}\,|\mathbf{X}|^{-\frac{\nu+p+1}{2}}\, e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{X}^{-1})}$$

Bayesians vs.Frequentists

You are no good when sample is small

You give a different answer for different priors