Recap

- **Bayes classifier** – assumes P_{XY} known, optimal for 0/1 loss

$$f(X) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} P(X = x|Y = y)P(Y = y)$$

Class conditional Distribution of features

Class distribution

- **Gaussian Bayes classifier** – assumes
  - Class distribution is Bernoulli/Multinomial
  - Class conditional distribution of features is Gaussian
- **Decision boundary** – (binary classification)

d-dim Gaussian Bayes classifier

$$f(X) = \arg\max_{Y=y} P(X = x|Y = y)P(Y = y)$$

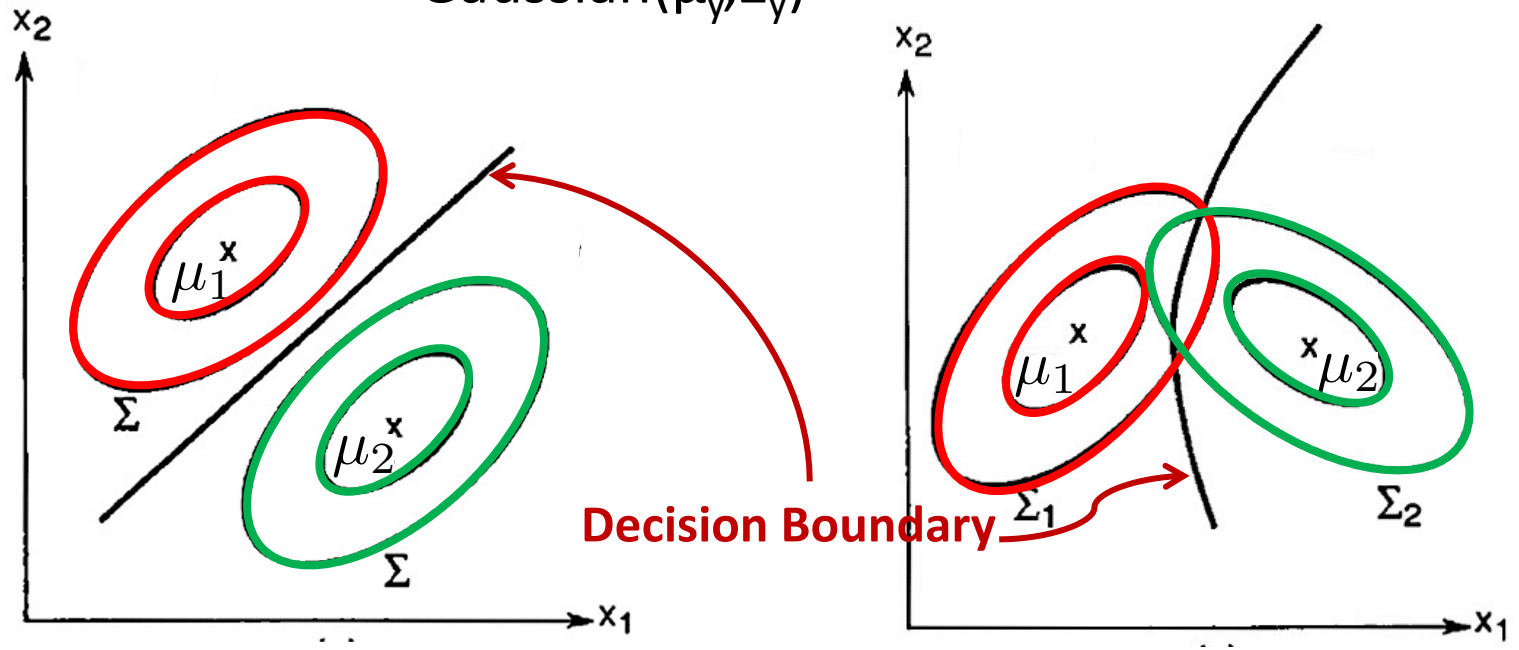
Class conditional Distribution of features

Class distribution

- What decision boundaries can we get in d-dim?

Gaussian(μ_y,Σ_y)

Bernoulli(θ)



Decision Boundary of Gaussian Bayes

- Decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

Compute the ratio

$$1 = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}$$

$$= \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \exp\left(-\frac{(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1)}{2} + \frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}\right) \frac{\theta}{1-\theta}$$

In general, this implies a quadratic equation in x. But if $\Sigma_1 = \Sigma_0$, then quadratic part cancels out and decision boundary is linear.

Glossary of Machine Learning

- Feature/Attribute
- iid
- Bayes classifier
- Class distribution
- Class conditional distribution of features
- Decision boundary

Maximum Likelihood Estimation (MLE)

Aarti Singh

Machine Learning 10-315
Sept 8, 2021

How to learn parameters from data? MLE

(Discrete case)

Learning parameters in distributions

$P(Y = \bullet) = \theta$ $P(Y = \bullet) = 1 - \theta$

Learning θ is equivalent to learning probability of head in coin flip.

- How do you learn that?

Data = 

Answer: 3/5

- Why??

Bernoulli distribution

Data, D = 

- Parameter θ : P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

Choose θ that maximizes the probability of observed data aka Likelihood

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

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

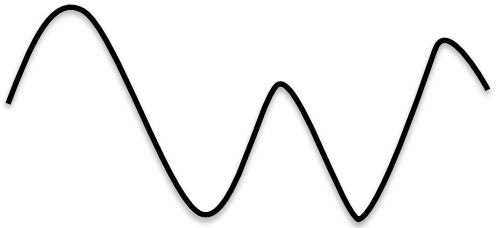
MLE of probability of head:

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5**

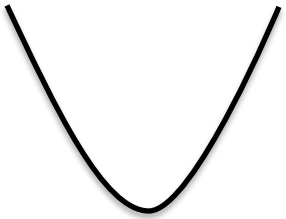
"Frequency of heads"

Short detour - Optimization

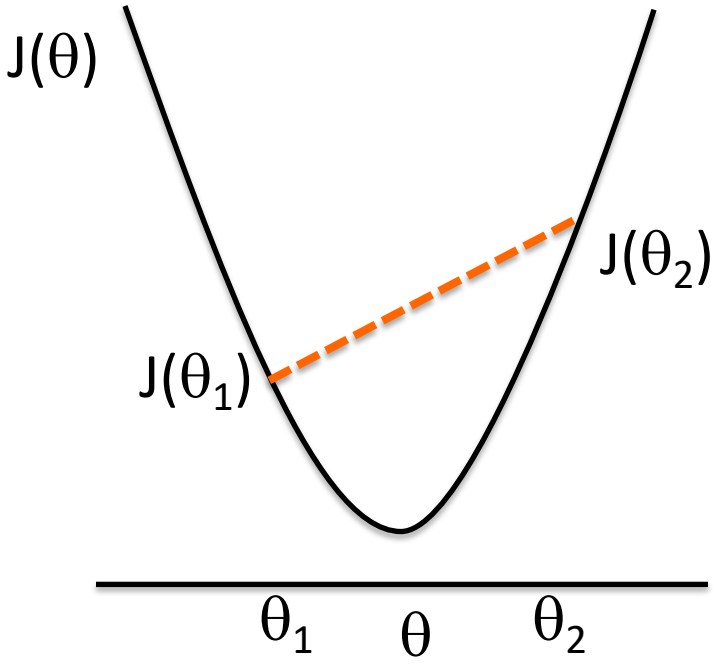
- Optimization objective $J(\theta)$
- Minimum value $J^* = \min_\theta J(\theta)$
- Minima (points at which minimum value is achieved) may not be unique



- If function is strictly convex, then minimum is unique

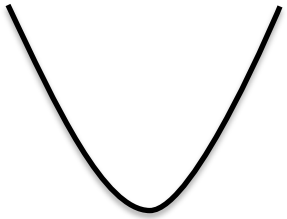


Convex functions

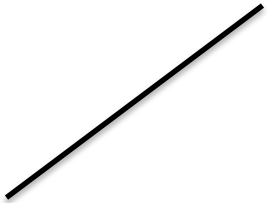


A function J(θ) is called **convex** if the line joining two points $J(\theta_1)$,$J(\theta_2)$ on the function does not go below the function on the interval $[\theta_1, \theta_2]$

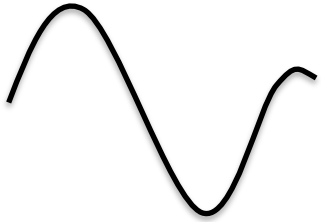
(Strictly) Convex functions have a unique minimum!



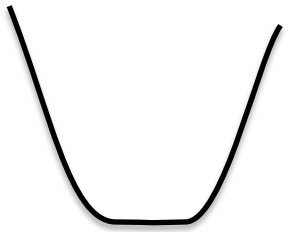
Convex



Both Concave & Convex



Neither



Convex but not strictly convex

Optimizing convex (concave) functions

- Derivative of a function
  - Partial derivative
- Derivative is zero at minimum of a convex function
- Second derivative is positive at minimum of a convex function

Optimizing convex (concave) functions

- What about
  - concave functions?
  - non-convex/non-concave functions?
  - functions that are not differentiable?
  - optimizing a function over a bounded domain aka constrained optimization?

Bernoulli MLE Derivation

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

Multinomial distribution

Data, D = rolls of a dice

- $P(1) = p_1$, $P(2) = p_2$, …, $P(6) = p_6$ $p_1+….+p_6 = 1$
- Rolls are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Multinomial(θ) distribution where

$$\theta = \{p_1, p_2, \dots, p_6\}$$

Choose θ that maximizes the probability of observed data aka "Likelihood"

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data

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

MLE of probability of rolls:

$$\hat{\theta}_{MLE} = \hat{p}_{1,MLE}, \ldots, \hat{p}_{6,MLE}$$

$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$

α_y ← Rolls that turn up y

$\sum_y \alpha_y$ ← Total number of rolls

"Frequency of roll y"

How to learn parameters from data? MLE

(Continuous case)

d-dim Gaussian distribution

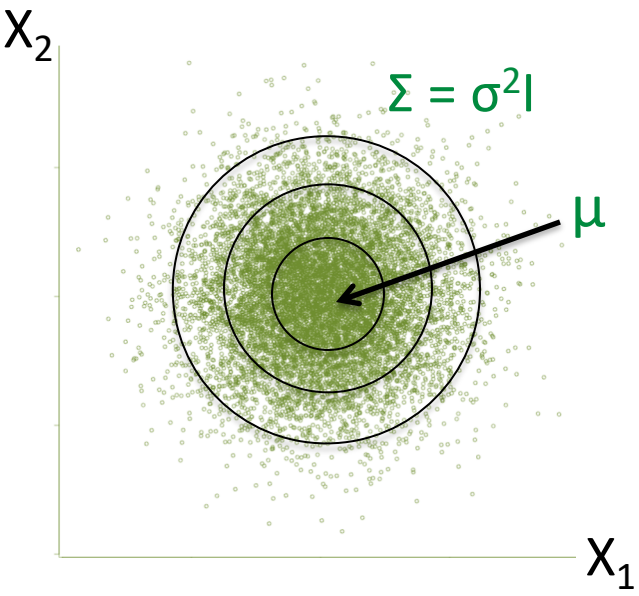
X is Gaussian N(μ, Σ)

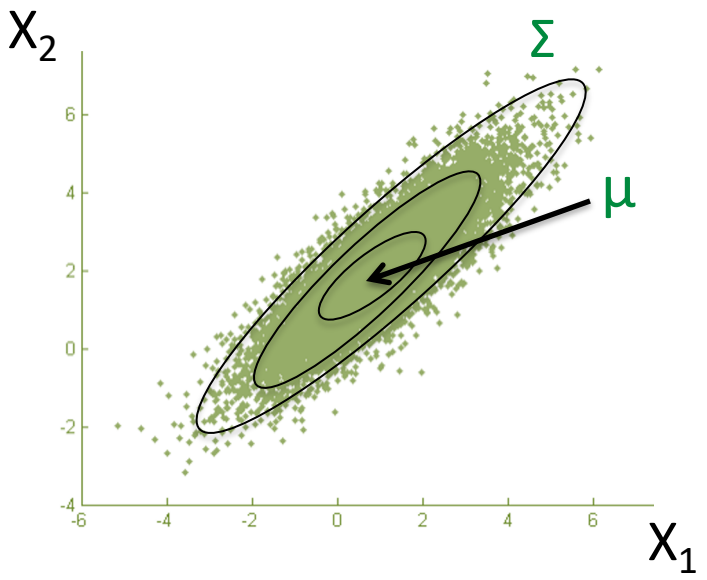
μ is d-dim vector, Σ is dxd dim matrix

$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

d=2

X = [X_1; X_2]



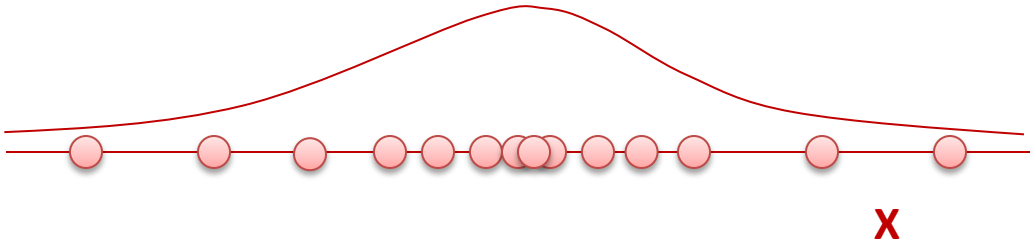


How to learn parameters from data? MLE

(Continuous case)

Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance

- Data are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Maximum Likelihood Estimation (MLE)

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

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

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

Independent draws

Maximum Likelihood Estimation (MLE)

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

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

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

Maximum Likelihood Estimation (MLE)

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

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

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

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

MLE for Gaussian mean

- Poll

$$P(D|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

A. $\max_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$

B. $\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$

C. $\max_{\mu} \mu^2 - 2\mu \sum_{i=1}^{n} X_i$

D. $\max_{\mu} n\mu^2 - 2\mu \sum_{i=1}^{n} X_i$

MLE for Gaussian mean and variance

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

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

Self exercise:

Derive MLE of variance?

d-dimensional versions?

MLE for uniform or exponential distribution?

More coming up in HW1

Unbiased estimates

- Bias of an estimate $E[\hat{\theta}] - \theta$
- Unbiased estimate if $E[\hat{\theta}] = \theta$
  - Is the MLE of mean unbiased?
  - Is the MLE of variance unbiased?
  - How can you make it unbiased?

Maximum A Posteriori (MAP) Estimation

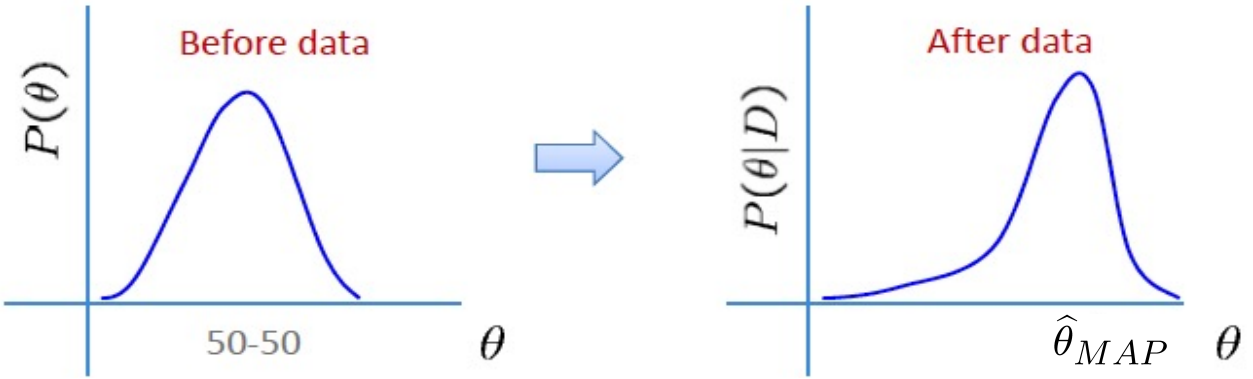
Aarti Singh

Machine Learning 10-315
Sept 13, 2021

Max A Posteriori (MAP) estimation

Can we bring in prior knowledge if data is not enough?

- Assume a prior (before seeing data D) distribution P(θ) for parameters θ



- Choose value that maximizes a posterior distribution P(θ|D) of parameters θ

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

How to choose prior distribution?

- $P(\theta)$
  - Prior knowledge about domain e.g. unbiased coin $P(\theta) = 1/2$
  - A mathematically convenient form e.g. "conjugate" prior

    If $P(\theta)$ is conjugate prior for $P(D|\theta)$, then Posterior has same form as prior

    Posterior = Likelihood x Prior

    $P(\theta|D) = P(D|\theta) \times P(\theta)$

e.g.	Beta	Bernoulli	Beta	θ = bias
	Gaussian	Gaussian	Gaussian	θ = mean μ (known Σ)
	inv-Wishart	Gaussian	inv-Wishart	θ = cov matrix Σ (known μ)

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta \mid D) \\ &= \arg\max_{\theta} P(D \mid \theta) P(\theta)\end{aligned}$$

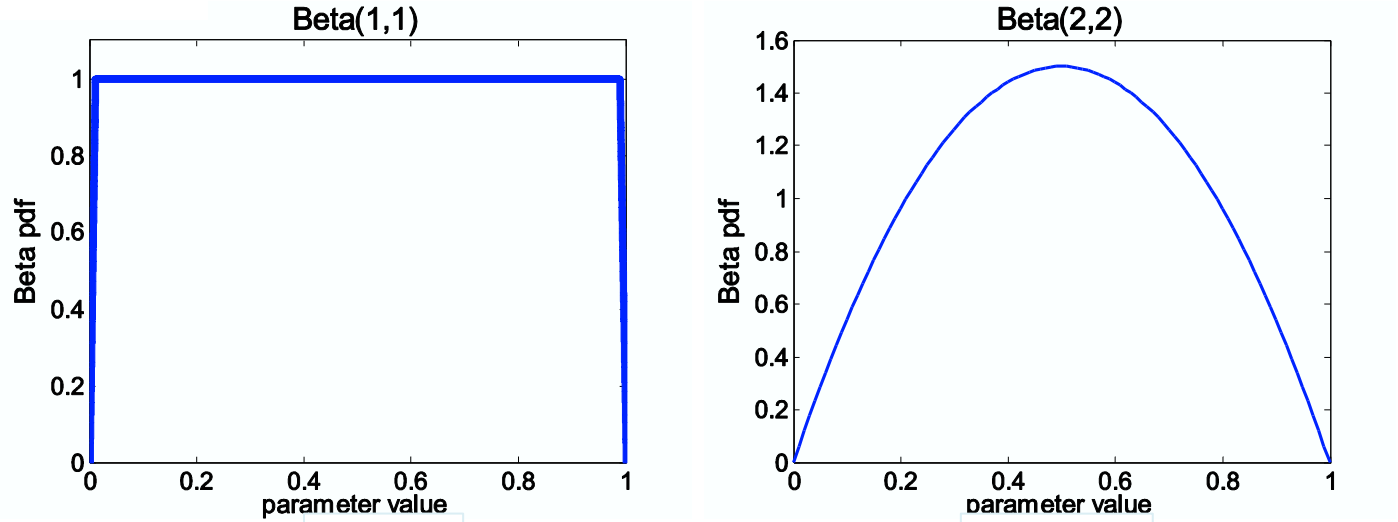
MAP estimate of probability of head (using Beta conjugate prior):

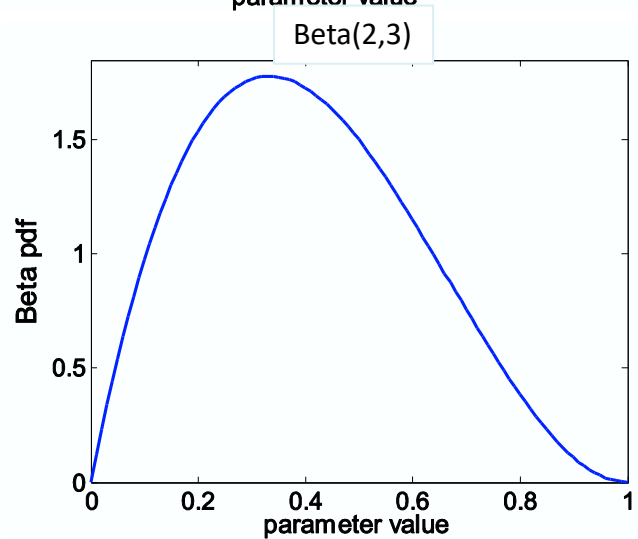
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

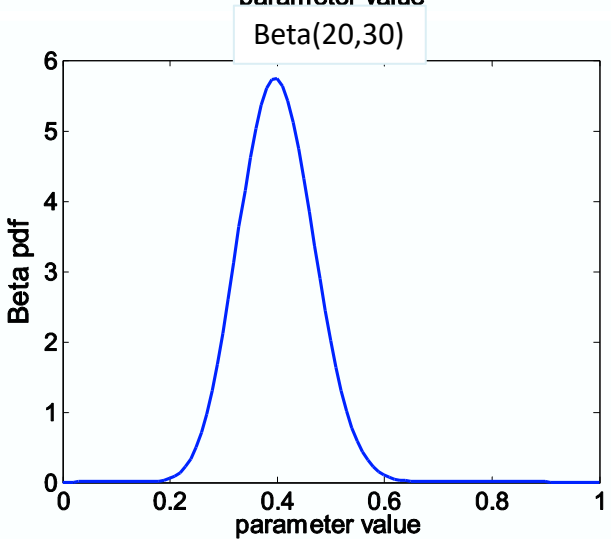
Beta distribution

$Beta(\beta_H, \beta_T)$

More concentrated as values of β_H, β_T increase







MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta \mid D) \\ &= \arg\max_{\theta} P(D \mid \theta) P(\theta)\end{aligned}$$

MAP estimate of probability of head (using Beta conjugate prior):

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

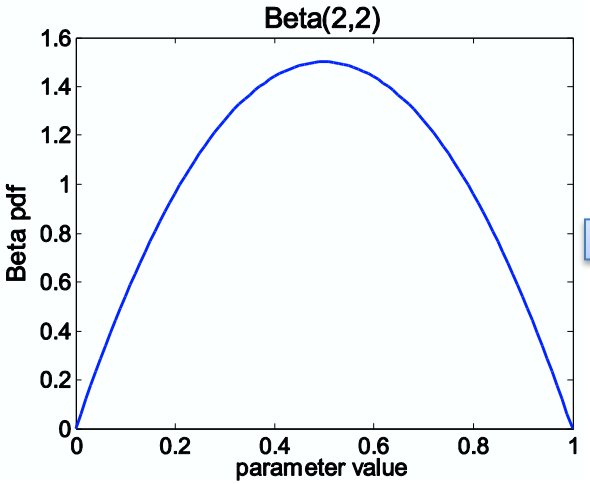
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

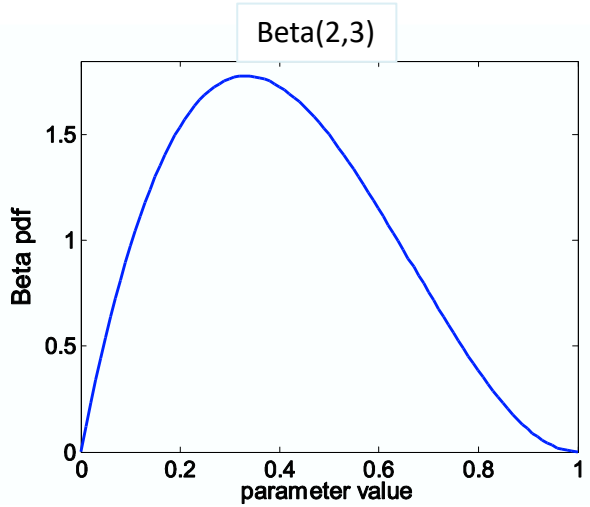
Count of H/T simply get added to parameters

Beta conjugate prior

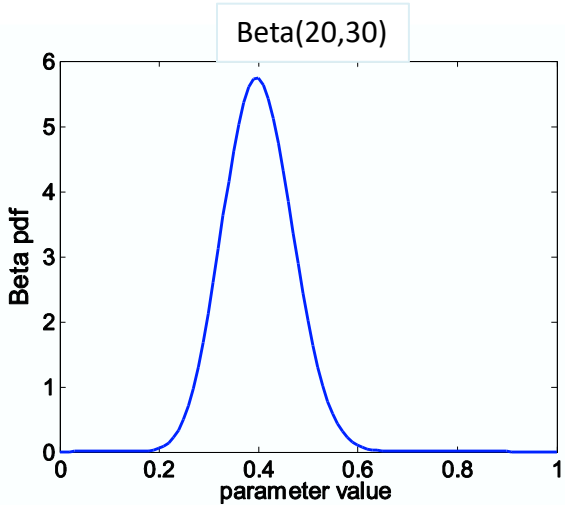
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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





After observing 1 Tail



After observing
18 Heads and
28 Tails

As $n = \alpha_H + \alpha_T$ increases, posterior distribution becomes more concentrated

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta \mid D) \\ &= \arg\max_{\theta} P(D \mid \theta) P(\theta)\end{aligned}$$

MAP estimate of probability of head:

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

Count of H/T simply get added to parameters

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

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

Equivalent to adding extra coin flips (β_H - 1 heads, β_T - 1 tails)

As we get more data, effect of prior is "washed out"

MAP estimation for Gaussian r.v.

Parameters $\theta = (\mu, \sigma^2)$

- Mean μ (known σ^2): Gaussian prior $P(\mu) = N(\eta, \lambda^2)$

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

$$\hat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}} \qquad \hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

As we get more data, effect of prior is "washed out"

- Variance σ^2 (known μ): inv-Wishart Distribution

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

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

- Maximum *a posteriori* (MAP) estimation

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

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

When is MAP same as MLE?