Maximum A Posteriori (MAP) Estimation

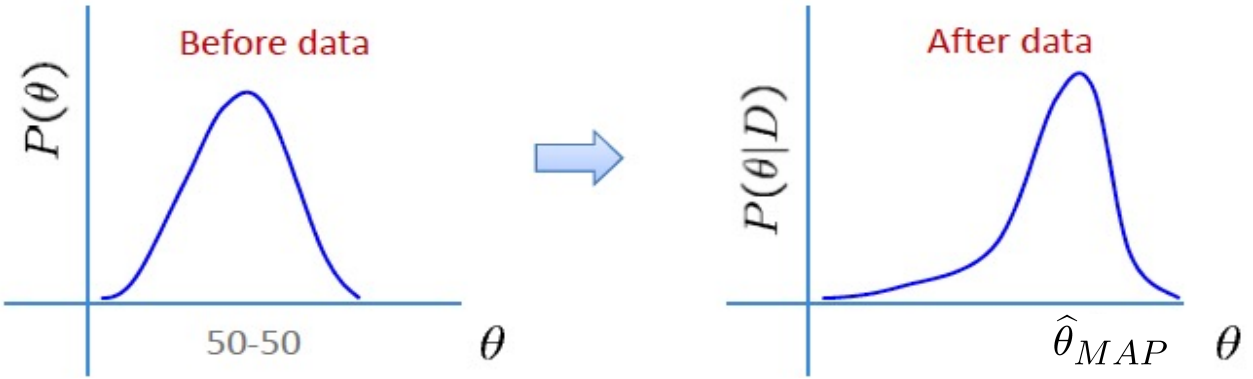
Aarti Singh

Machine Learning 10-315
Jan 31, 2022

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

$$\text{Posterior} = \text{Likelihood} \times \text{Prior}$$
$$P(\theta|D) = P(D|\theta) \times P(\theta)$$

e.g.	Beta	Bernoulli	Beta	θ = bias of coin
	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

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

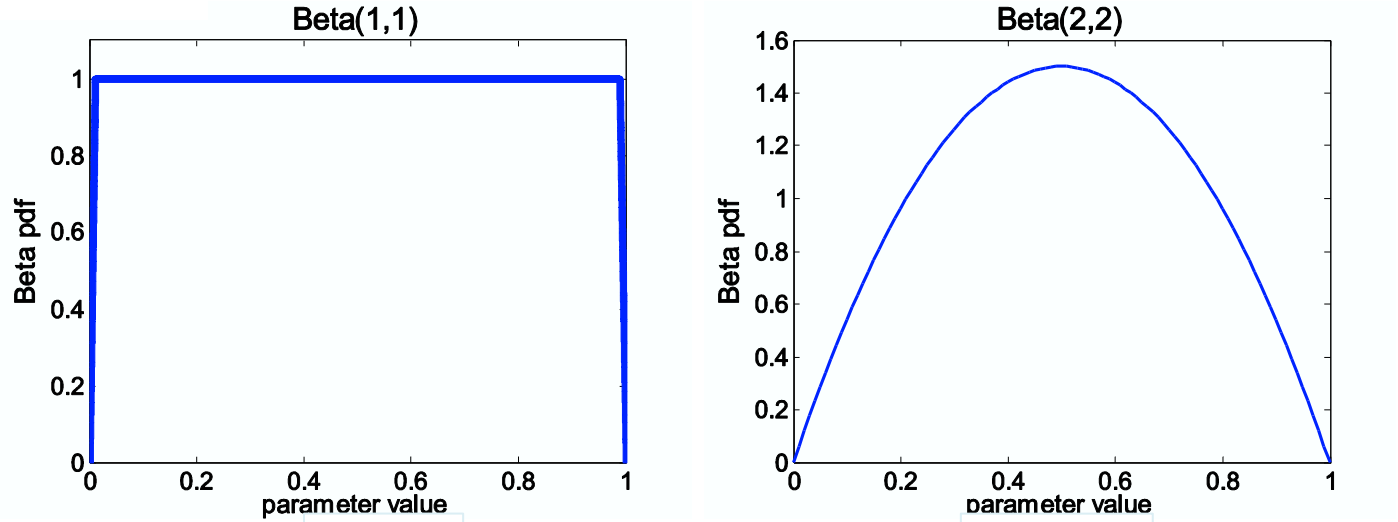
MAP estimate of probability of head (using Beta conjugate prior):

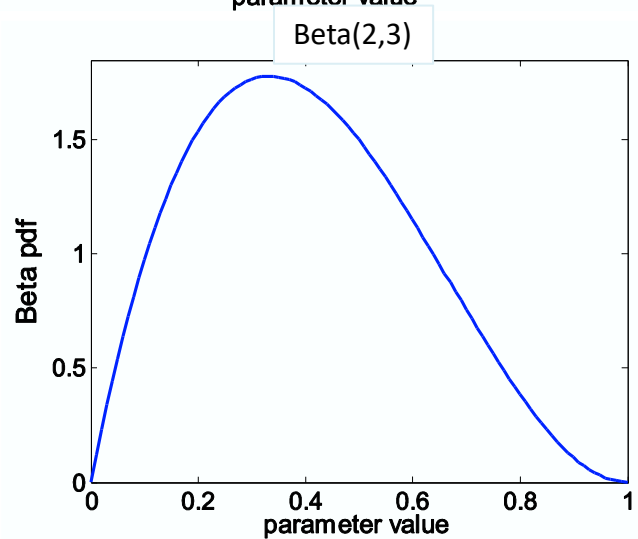
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

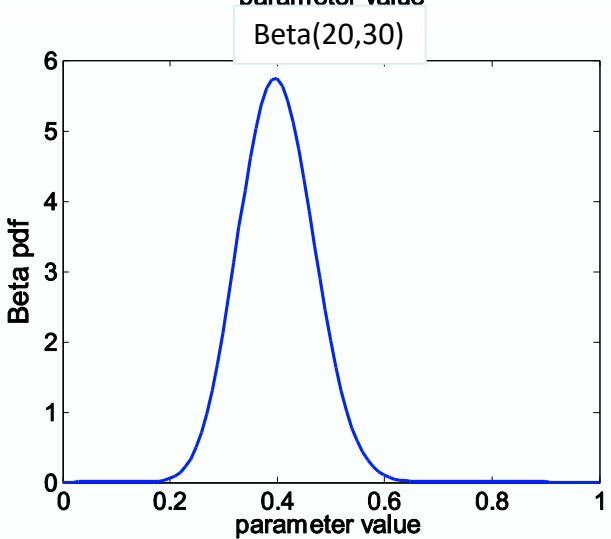
Beta distribution

$Beta(\beta_H, \beta_T)$

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):

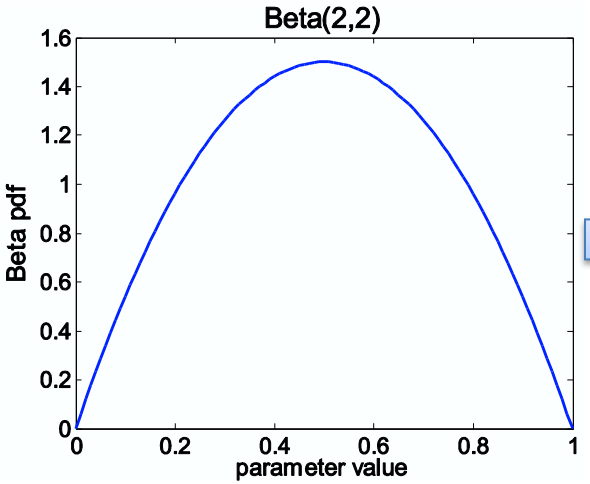
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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

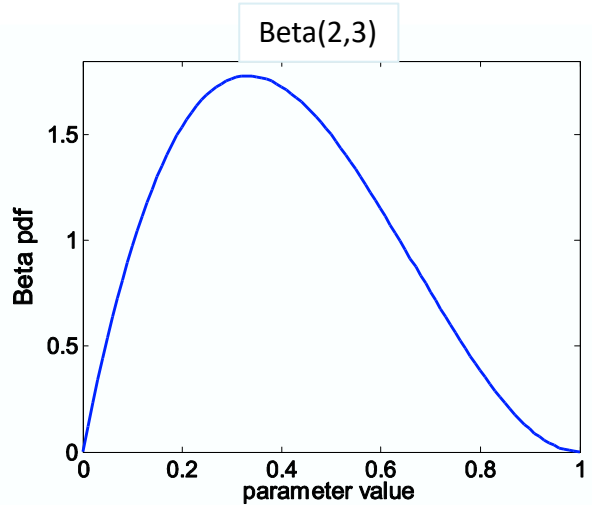
Count of H/T simply get added to parameters

Beta conjugate prior

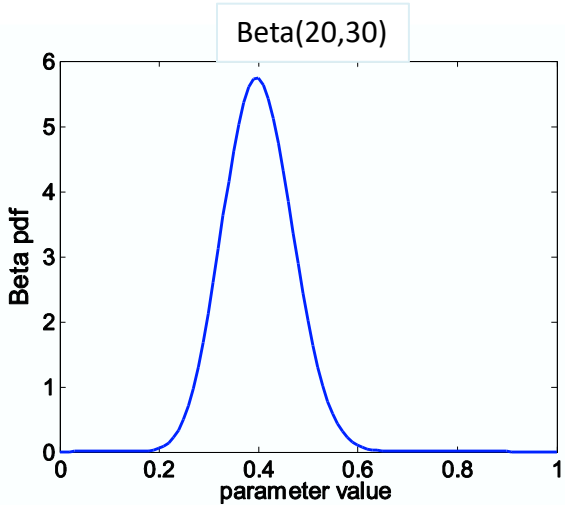
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$



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



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

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

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

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

- Maximum *a posteriori* (MAP) estimation

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

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

Poll

- When is MAP same as MLE?

A. When posterior is same as prior

B. When prior is uniform

C. When prior is zero for all values except one value of θ