Parametric Models: Prior Information, MAP

Manuela Veloso
Co-instructor: Pradeep Ravikumar

Thanks to past instructors

A.Moore tutorials www.cs.cmu.edu/~awm/tutorials

Machine Learning

Jan 24, 2018

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Recall: Your first consulting job

A billionaire from the suburbs of Seattle asks you a question:

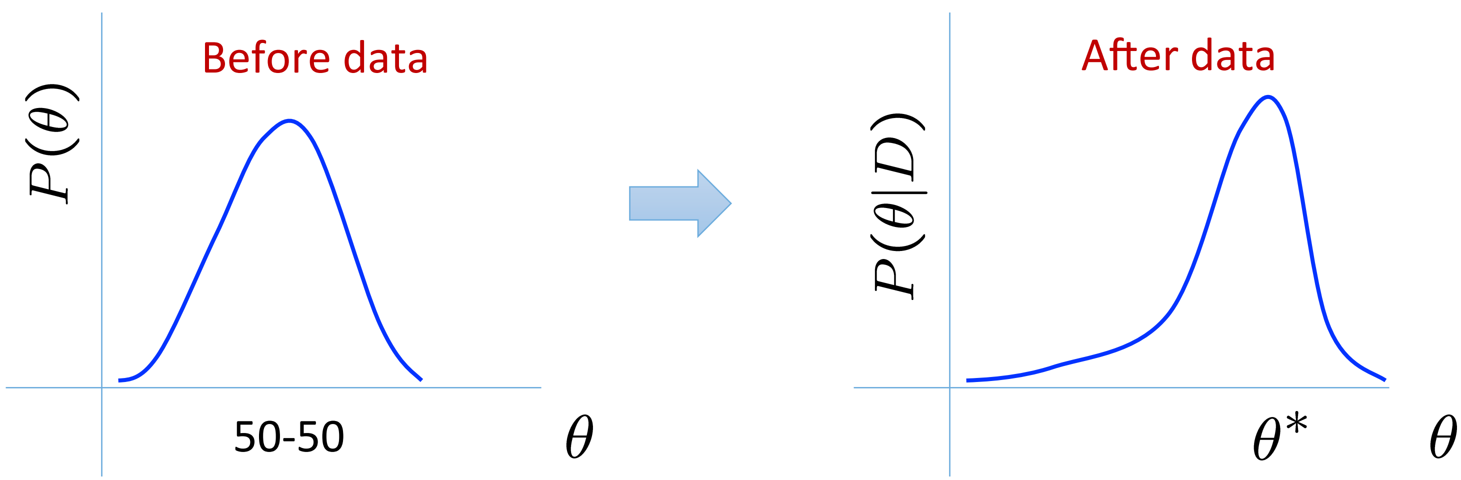
- He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
- You say: Please flip it a few times:

- You say: The probability is: **3/5** because… frequency of heads in all flips
- **He says: But can I put money on this estimate?**
- You say: ummm…. Maybe not.
  - Not enough flips (less than sample complexity)

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

- Use the *prior* formation; Estimate a *distribution over possible values of* θ*, given the data*



Bayesian Learning

Use Bayes rule:

$$P(\underset{\text{Parameters}}{\theta} \mid \underset{\text{Data}}{\mathcal{D}}) = \frac{\overset{\text{likelihood}}{P(\mathcal{D} \mid \theta)}\,\overset{\text{prior}}{P(\theta)}}{P(\mathcal{D})}$$

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**

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

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**

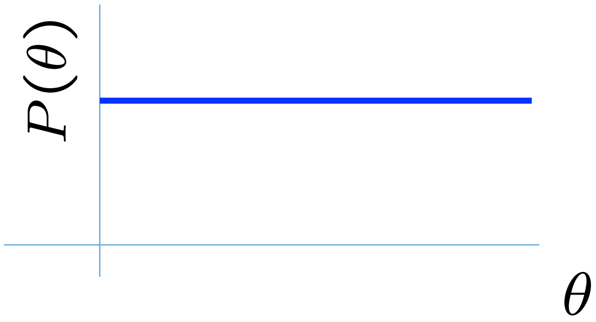
Prior distribution

From where do we get the prior?

- Represents expert knowledge (philosophical approach)
- Simple posterior form (engineer's approach)

Uninformative priors:

- Uniform distribution



General prior: computational issues for online learning

Conjugate Prior

Consider a *family* of probability distributions characterized by some parameter θ (possibly a single number, possibly a tuple).

A prior is a *conjugate prior* if:

- If it is a member of this family;
- and if all possible posterior distributions are also members of this family.

$P(\theta)$ and $P(\theta|D)$ have the same form as a function of θ.

Closed-form representation of posterior!

Conjugate Prior
(Can check table of conjugate prior distributions.)

- P(θ) and P(θ|D) have the same form as a function of theta

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

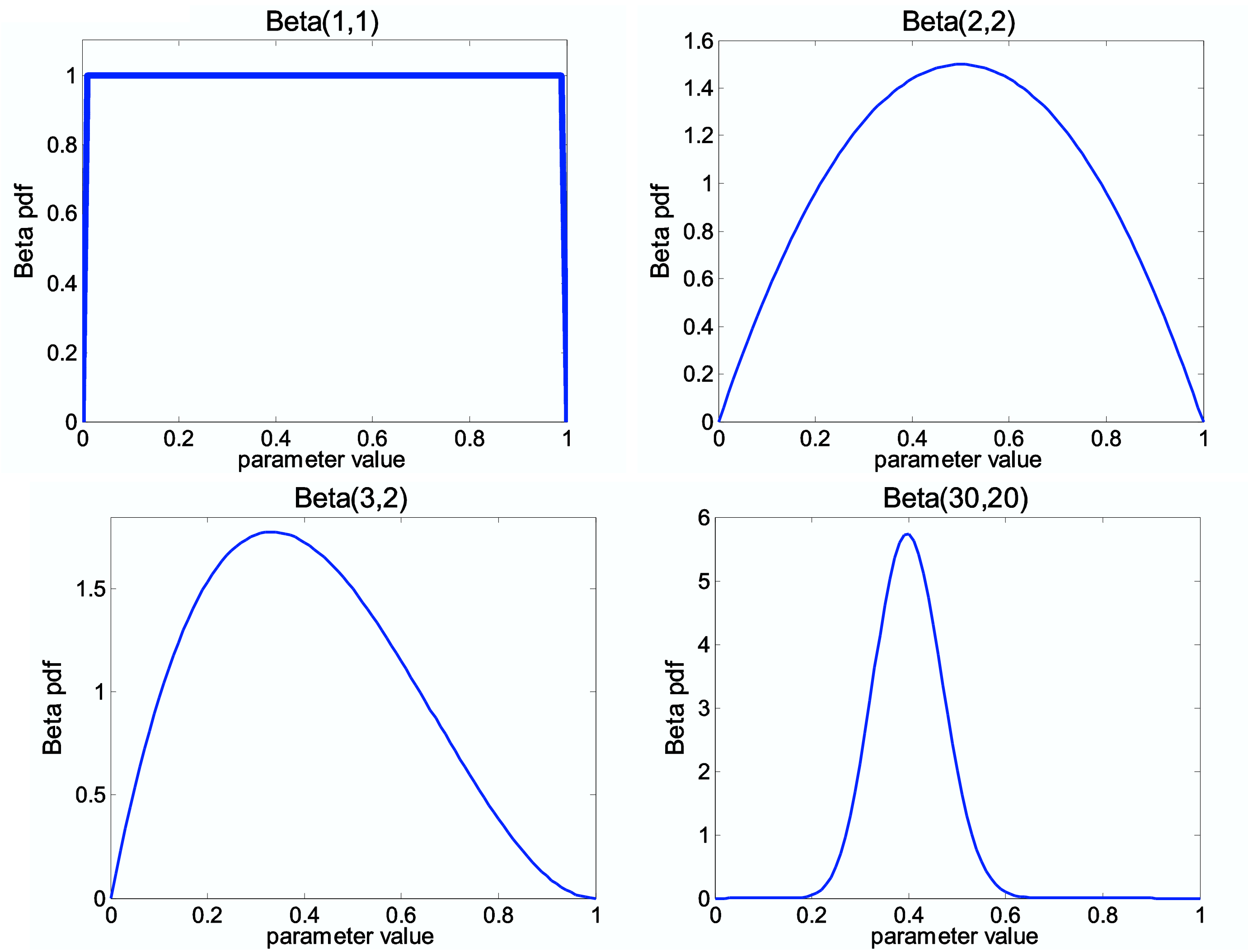
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

Beta distribution

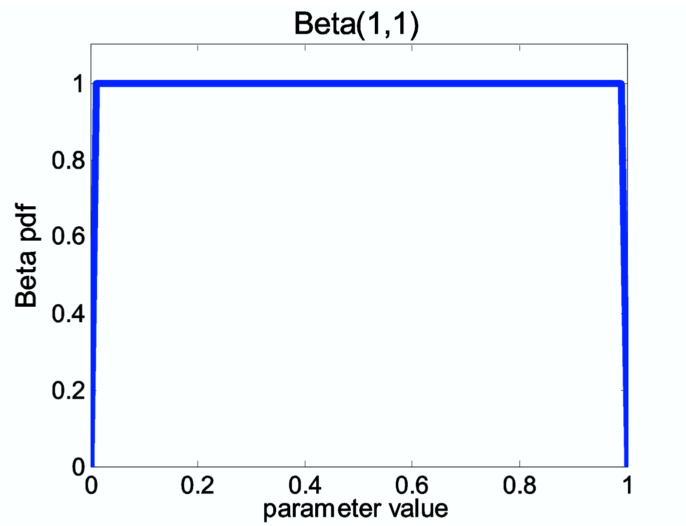
$Beta(\beta_H, \beta_T)$

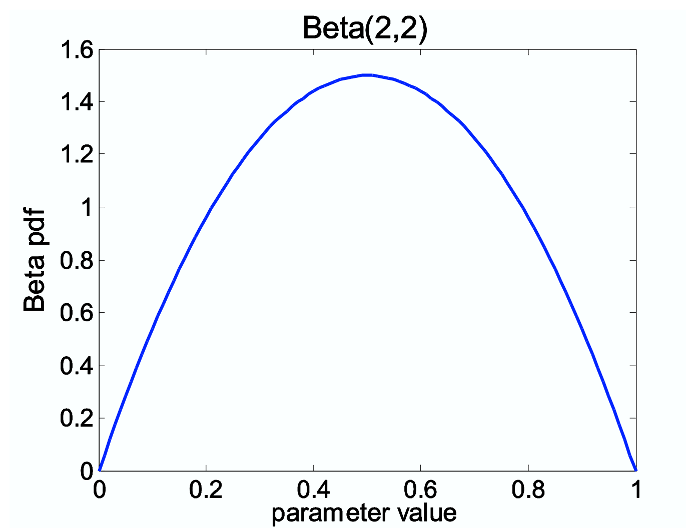
More concentrated as values of β_H, β_T increase



think the coin is fair – it is “close” to 50-50

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

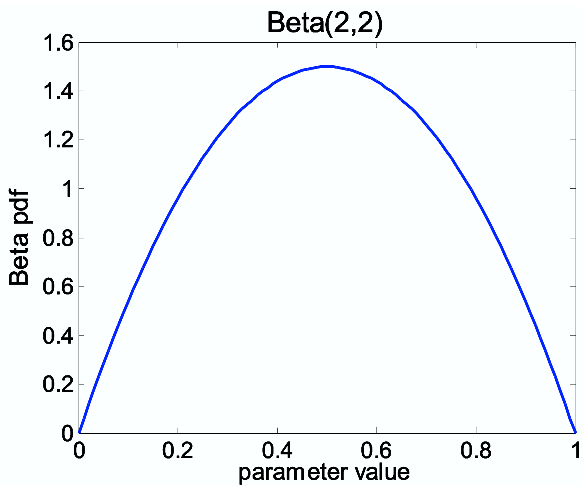


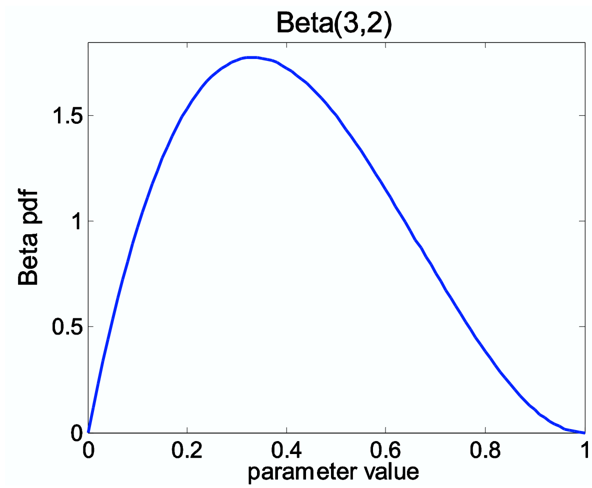


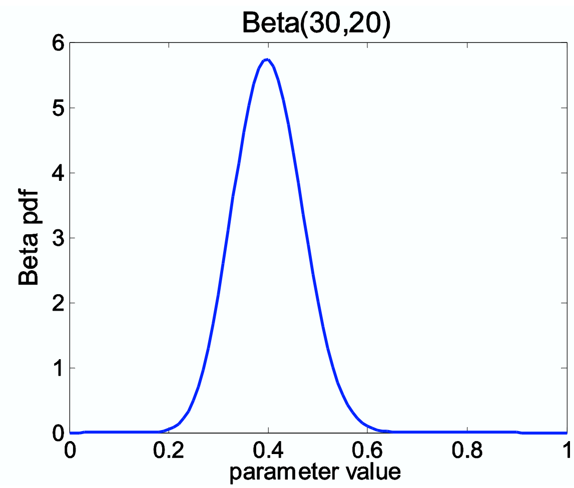
....

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”

Conjugate Prior

- $P(\theta)$ and $P(\theta|D)$ have the same form

Eg. 2 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.

Posterior Distribution

- The approach seen so far is what is known as a **Bayesian** approach
- Prior information encoded as a **distribution** over possible values of parameter
- Using the Bayes rule, you get an updated **posterior** distribution over parameters, which you provide with flourish to the Billionaire
- But the billionaire is not impressed
  - Distribution? I just asked for one number: is it 3/5, 1/2, what is it?
  - How do we go from a distribution over parameters, to a single estimate of the true parameters?

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

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

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

MAP estimate of probability of head:

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

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

Mode of Beta distribution

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?

MLE vs. MAP

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

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! ...with prob 1 ☺

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

- Beta prior equivalent to extra coin flips
- As $n \rightarrow$ infty, prior is "forgotten"
- **But, for small sample size, prior is important!**

MLE vs MAP

You are no good when sample is small

You give a different answer for different priors

MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

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

MAP for Gaussian Mean

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

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

MAP of Gaussian variance - Later

Prior Information

- In the Bayesian approach, the prior information is encoded through a prior distribution over the parameters
- Seems onerous: the distribution typically seems to be obtained from convenience (conjugate distribution)
- What other ways can we encode our prior knowledge about the parameters?
- A non-Bayesian approach is via constraints: later

Autonomous Robot Navigation

Mobile robots have sensors and motors

How do they move?

- Need to know *where* they are
- Combine apriori knowledge with data
- Not learning, but computing ~MAP

Apriori: own motion, and sensing

Discussion

MAP can be seen as “superior” to MLE

- Use of priors
- Good estimates from few data

Robustness tradeoff

- What if the prior is wrong?

Summary

- Conditional probabilities
- Bayes Rule
- Priors, conjugate prior
- MAP - maximum aposteriori estimate
- MLE and MAP
- Example: Bayesian update for robot localization estimate