



CMSC726 Spring 2006: Parameter Estimation aka Statistical Learning 101

readings: Parameter Estimation &
Foundations Handouts

sources: course slides are based on
material from a variety of sources,
including Tom Dietterich, **Carlos
Guestrin**, Rich Maclin, Ray Mooney,
Andrew Moore, Andrew Ng, Jude



Your first consulting job

- A highroller from Las Vegas asks you a question:
 - He says: I have thumbtack, if I flip it, what's the probability it will fall with the flat side up?
 - You say: Please flip it a few times:
 - You say: The probability is:
 - **He says: Why???**
 - You say: Because...

● ● ● | Thumbtack – Binomial Distribution

- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$
- Flips are i.i.d.:
 - Independent events
 - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

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

● ● ● | Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
 - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \ln P(\mathcal{D} | \theta)\end{aligned}$$

● ● ● | 'Learning' algorithm

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0$

● ● ● | How many flips do I need?

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

- Highroller says: I flipped 3 heads and 2 tails.
- You say: $\theta = 3/5$, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Humm... The more the merrier???

● ● ● | Simple bound
(based on Hoeffding's inequality)

○ For $N = \alpha_H + \alpha_T$, and $\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

○ Let θ^* be the true parameter, for any $\epsilon > 0$:

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

● ● ● | PAC Learning

- PAC: Probably Approximate Correct
- Highroller says: I want to know the thumbtack parameter θ , within $\epsilon = 0.1$, with probability at least $1 - \delta = 0.95$. How many flips?

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

● ● ● | What about prior

- Highroller says: Wait, I know that the thumbtack is “close” to 50-50. What can you say?
- **You say: I can learn it the Bayesian way...**
- Rather than estimating a single θ , we obtain a distribution over possible values of θ

● ● ● | Bayesian Learning

- Use Bayes rule:

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

- Or equivalently:

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

● ● ● | Bayesian Learning for Thumbtack

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

- Likelihood function is simply Binomial:

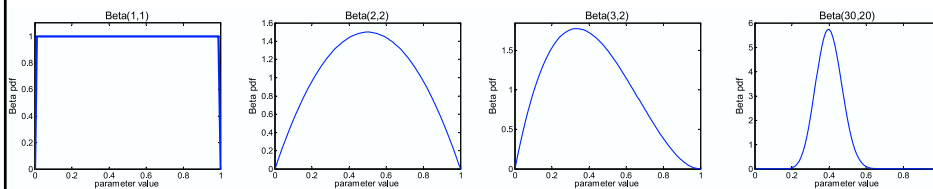
$$P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$$

- What about prior?
 - Represent expert knowledge
 - Simple posterior form
- Conjugate priors:
 - Closed-form representation of posterior
 - **For Binomial, conjugate prior is Beta distribution**

● ● ● | Beta prior distribution – P(θ)

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

$$\gamma = \frac{\Gamma(\beta_H + \beta_T)}{\Gamma(\beta_H)\Gamma(\beta_T)}$$

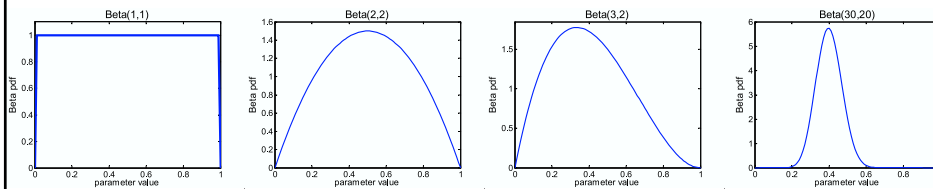


- Likelihood function: $P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

● ● ● | Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

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



● ● ● | Using Bayesian posterior

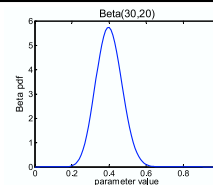
- Posterior distribution:

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

- Bayesian inference:
 - No longer single parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta | \mathcal{D}) d\theta$$

- Integral is often hard to compute



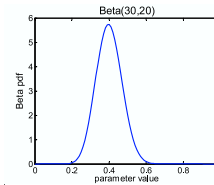
● ● ● | MAP: Maximum a posteriori approximation

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

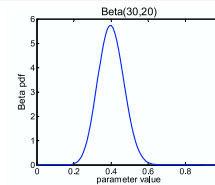
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta | \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain
- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D}) \quad E[f(\theta)] \approx f(\hat{\theta})$$



● ● ● | MAP for Beta distribution



$$P(\theta | \mathcal{D}) = \gamma' \theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$\gamma' = \frac{\Gamma(\alpha_H + \beta_H + \alpha_T + \beta_T)}{\Gamma(\alpha_H + \beta_H) \Gamma(\alpha_T + \beta_T)}$$

- MAP: use most likely parameter:

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

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

● ● ● | Summary

- Parameter estimation 101:
 - MLE
 - Bayesian estimation
 - MAP