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 \( \alpha_H \) Heads and \( \alpha_T \) Tails
  \[
P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
  \]

Maximum Likelihood Estimation

- **Data:** Observed set \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails
- **Hypothesis:** Binomial distribution
- Learning \( \theta \) is an optimization problem
  - What’s the objective function?
- MLE: Choose \( \hat{\theta} \) that maximizes the probability of observed data:
  \[
  \hat{\theta} = \arg \max_{\theta} \ P(D \mid \theta) = \arg \max_{\theta} \ \ln P(D \mid \theta)
  \]
| ‘Learning’ algorithm

\[
\hat{\theta} = \arg\max_{\theta} \ln P(D | \theta) = \arg\max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
\]

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

| 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 $\theta^*$ be the true parameter, for any $\epsilon > 0$:

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

PAC Learning

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

$$P(\mid \hat{\theta} - \theta^* \mid \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$
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 $\theta$, we obtain a distribution over possible values of $\theta$

---

Bayesian Learning

- Use Bayes rule:
  \[
  P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}
  \]

- Or equivalently:
  \[
  P(\theta \mid D) \propto P(D \mid \theta)P(\theta)
  \]
Bayesian Learning for Thumbtack

\[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]

- Likelihood function is simply Binomial:
  \[ P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \]

- 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(\theta)

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

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

- Likelihood function:
  \[ P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \]
- Posterior:
  \[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]
Posterior distribution

- Prior: $\text{Beta}(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:

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

Using Bayesian posterior

- Posterior distribution:

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

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

- Integral is often hard to compute
MAP: Maximum a posteriori approximation

\[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

\[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta \]

- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

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

### MAP for Beta distribution

\[ P(\theta \mid D) = \gamma^' \theta^{\alpha_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_H - 1} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_H) \]

\[ \gamma^' = \frac{\Gamma(\alpha_H + \beta_H + \alpha_T + \beta_T)}{\Gamma(\alpha_H + \beta_H) \Gamma(\alpha_T + \beta_T)} \]

- MAP: use most likely parameter:

\[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid D) = \]

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

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