EE 511
Point Estimation

Instructor: Hanna Hajishirzi
hannaneh@washington.edu

Recap

• Learning is
  – Collect some data
  – Choose a hypothesis class or model
  – Choose a loss function:
  – Choose an optimization procedure
  – Justifying the accuracy of the estimate
Your first consulting job

• A billionaire from the suburbs of Seattle asks you a question:
  – **He says:** I have thumbtack, if I flip it, what’s the probability it will fall with the nail up?
  – **You say:** Please flip it a few times:

![Thumbtacks](image)

  – **You say:** The probability is:
    • \( P(H) = \frac{3}{5} \)

  – **He says:** Why???
  – **You say:** Because...
Thumbtack – Binomial Distribution

• $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1-\theta$

• Flips are $i.i.d.$: $D = \{x_i| i=1...n\}$, $P(D | \theta) = \prod_i P(x_i | \theta)$
  − Independent events
  − Identically distributed according to Binomial distribution

• Sequence $D$ of $\alpha_H$ Heads and $\alpha_T$ Tails

\[
P(D | \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}
\]
Maximum Likelihood Estimation

• **Data:** Observed set $D$ of $\alpha_H$ Heads and $\alpha_T$ Tails

• **Hypothesis space:** Binomial distributions

• **Learning:** finding $\theta$ is an optimization problem
  - What’s the objective function?

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

• **MLE:** Choose $\theta$ to maximize probability of $D$

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

\[
= \arg \max_{\theta} \ln P(D \mid \theta)
\]
Your first parameter learning algorithm

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

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

• Set derivative to zero, and solve!

\[ \frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}] \]

\[ = \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1 - \theta)] \]

\[ = \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1 - \theta) \]

\[ = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0 \]

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]
But, how many flips do I need?

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

- Billionaire 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: Umm... The more the merrier???
- He says: Is this why I am paying you the big bucks???
A bound (from Hoeffding’s inequality)

• For $N = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

• Let $\theta^*$ be the true parameter, for any $\varepsilon > 0$:

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

• **PAC**: Probably Approximately Correct
• **Billionaire says**: I want to know the thumbtack $\theta$, within $\varepsilon = 0.1$, with probability at least $1 - \delta = 0.95$.
• **How many flips? Or, how big do I set $N$?**

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

$$\delta \geq 2e^{-2N\varepsilon^2} \geq P(\text{mistake})$$

$$\ln \delta \geq \ln 2 - 2N\varepsilon^2$$

$$N \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$

Interesting! Let's look at some numbers!

• $\varepsilon = 0.1$, $\delta=0.05$

$$N \geq \frac{\ln(2/0.05)}{2 \times 0.1^2} \approx \frac{3.8}{0.02} = 190$$
MLE Recap

• Learning is
  – Collect some data
    • E.g., coin flips
  – Choose a hypothesis class or model
    • E.g., Binomial
  – Choose a loss function:
    • E.g., data likelihood
  – Choose an optimization procedure
    • Set derivatives to zero to obtain MLE
  – Justifying the accuracy of the estimate
    • E.g., Hoeffding’s inequality
What if I have prior beliefs?

• Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
• 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:

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

- Or equivalently:

\[ P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta) \]

- Also, for uniform priors:

\[ P(\theta) \propto 1 \]

\[ P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta) \]

\[ \rightarrow \text{reduces to MLE objective} \]
Bayesian Learning for Thumbtacks

\[ P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta) \]

Likelihood function is 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 – $P(\theta)$

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

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

\[ P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \theta^{\beta_H-1}(1-\theta)^{\beta_T-1} \]

\[ = \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1} = Beta(\alpha_H+\beta_H, \alpha_T+\beta_T) \]
Posterior distribution

- **Prior**: \( Beta(\beta_H, \beta_T) \)
- **Data**: \( \alpha_H \) heads and \( \alpha_T \) tails
- **Posterior distribution**:

\[
P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)
\]
MAP for Beta distribution

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:
  \[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
What about continuous variables?

• Billionaire says: If I am measuring a continuous variable, what can you do for me?

• You say: Let me tell you about Gaussians...

\[ P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Some properties of Gaussians

• Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  
  \( X \sim N(\mu, \sigma^2) \)
  
  \( Y = aX + b \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2) \)

• Sum of Gaussians is Gaussian
  
  \( X \sim N(\mu_X, \sigma^2_X) \)
  
  \( Y \sim N(\mu_Y, \sigma^2_Y) \)
  
  \( Z = X + Y \Rightarrow Z \sim N(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y) \)

• Easy to differentiate, as we will see soon!
Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: $\mu$
  - Variance: $\sigma$

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

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>Exam Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>99</td>
<td>89</td>
</tr>
</tbody>
</table>
MLE for Gaussian: \[ P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- Prob. of i.i.d. samples \( D = \{x_1, \ldots, x_N\} \):

\[
P(D \mid \mu, \sigma) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}
\]

\[ \mu_{MLE}, \sigma_{MLE} = \arg \max_{\mu, \sigma} P(D \mid \mu, \sigma) \]

- Log-likelihood of data:

\[
\ln P(D \mid \mu, \sigma) = \ln \left[ \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\
= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}
\]
Your second learning algorithm: MLE for mean of a Gaussian

• What’s MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} | \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\mu} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= - \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0$$

$$= - \sum_{i=1}^{N} x_i + N\mu = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$
MLE for variance

- Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(D \mid \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^3} = 0
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Learning Gaussian parameters

• MLE:

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

• BTW. MLE for the variance of a Gaussian is **biased**
  – Expected result of estimation is **not** true parameter!
  – Unbiased variance estimator:

\[
\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Bayesian learning of Gaussian parameters

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

- Prior for mean:

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

• Learning is
  – Collect some data
    • E.g., coin flips
  – Choose a hypothesis class or model
    • E.g., Binomial and prior based on expert knowledge
  – Choose a loss function:
    • E.g., parameter posterior likelihood
  – Choose an optimization procedure
    • Set derivatives to zero to obtain MAP
  – Justifying the accuracy of the estimate
    • E.g., if the model is correct your are doing best possible