CSE546: Point Estimation
Winter 2012

Luke Zettlemoyer

Slides adapted from Carlos Guestrin and Dan Klein
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](image1.png)

  – **You say:** The probability is:
    • \( P(H) = 3/5 \)
  – **He says:** Why???
  – **You say:** Because...
Random Variables

• A random variable is some aspect of the world about which we (may) have uncertainty
  – R = Is it raining?
  – D = How long will it take to drive to work?
  – L = Where am I?

• We denote random variables with capital letters

• Random variables have domains
  – R in \{true, false\} (sometimes write as \{+r, ¬r\})
  – D in \[0, \infty\)
  – L in possible locations, maybe \{(0,0), (0,1), \ldots\}
Probability Distributions

- Discrete random variables have distributions

\[ P(T) \]

<table>
<thead>
<tr>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>0.5</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[ P(W) \]

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>fog</td>
<td>0.3</td>
</tr>
<tr>
<td>meteor</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- A discrete distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number

\[ P(W = rain) = 0.1 \quad P(rain) = 0.1 \]

- Must have:

\[ \forall x \ P(x) \geq 0 \quad \sum_x P(x) = 1 \]
Joint Distributions

- A **joint distribution** over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or outcome):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) \quad \text{and} \quad P(x_1, x_2, \ldots x_n)$$

- Size of distribution if $n$ variables with domain sizes $d$?

- Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1,x_2,\ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

- For all but the smallest distributions, impractical to write out

<table>
<thead>
<tr>
<th>$T$</th>
<th>$W$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

\[
P(T, W) = \begin{array}{ccc}
T & W & P \\
hot & sun & 0.4 \\
hot & rain & 0.1 \\
cold & sun & 0.2 \\
cold & rain & 0.3 \\
\end{array}
\]

\[
P(T) = \sum_s P(t, s)
\]

\[
P(W) = \sum_t P(t, s)
\]

\[
P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)
\]
Conditional Probabilities

• A simple relation between joint and conditional probabilities
  – In fact, this is taken as the definition of a conditional probability

\[ P(a|b) = \frac{P(a, b)}{P(b)} \]

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[ P(W = r|T = c) = ??? \]
Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others.

| Condition | P(W|T)          | Joint Distribution |
|-----------|----------------|-------------------|
|           | P(W|T = hot)    | P(T, W)           |
|           | W     | P    | T     | W     | P     |
| sun      | 0.8   |      | hot   | sun   | 0.4   |
| rain     | 0.2   |      | hot   | rain  | 0.1   |
|           | P(W|T = cold)  |                   |
|           | W     | P    | cold  | sun   | 0.2   |
|           | rain  | 0.6  | cold  | rain  | 0.3   |
The Product Rule

- Sometimes have conditional distributions but want the joint

\[ P(x|y) = \frac{P(x, y)}{P(y)} \leftrightarrow P(x, y) = P(x|y)P(y) \]

- Example:

| $P(W)$ | $P(D|W)$ | $P(D, W)$ |
|--------|-----------|-----------|
|        | D         | W         | P         |        | D         | W         | P         |
| R      | wet       | sun       | 0.1       | wet     | sun       | 0.08      |
| sun    | dry       | sun       | 0.9       | dry     | sun       | 0.72      |
| rain   | wet       | rain      | 0.7       | wet     | rain      | 0.14      |
|        | dry       | rain      | 0.3       | dry     | rain      | 0.06      |
Bayes’ Rule

• Two ways to factor a joint distribution over two variables:

\[ P(x, y) = P(x|y)P(y) = P(y|x)P(x) \]

• Dividing, we get:

\[ P(x|y) = \frac{P(y|x)}{P(y)} P(x) \]

• Why is this at all helpful?
  – Lets us build one conditional from its reverse
  – Often one conditional is tricky but the other one is simple
  – Foundation of many practical systems (e.g. ASR, MT)

• In the running for most important ML equation!
Thumbtack – Binomial Distribution

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

- Flips are \textit{i.i.d.}: \( D = \{ x_i \mid i = 1 \ldots n \} \), \( P(D \mid \theta) = \Pi_i P(x_i \mid \theta) \)
  - 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:** 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

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

- Set derivative to zero, and solve!

\[ \frac{d}{d\theta} \ln P(D \mid \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}] \]
\[ = \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1 - \theta)] \]
\[ = \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1 - \theta) \]
\[ = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0 \]
\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]
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 $\epsilon = 0.1$, with probability at least $1 - \delta = 0.95$.
• **How many flips?** Or, how big do I set $N$?

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

\[
\delta \geq 2e^{-2N\epsilon^2} \geq P(\text{mistake})
\]

\[
\ln \delta \geq \ln 2 - 2N\epsilon^2
\]

\[
N \geq \frac{\ln(2/\delta)}{2\epsilon^2}
\]

Interesting! Let's look at some numbers!

• $\epsilon = 0.1$, $\delta=0.05$

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

---

![Graphs showing the change in belief distribution before and after observing flips.](image-url)
Bayesian Learning

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

- Or equivalently: 
  \[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]

- Also, for uniform priors: 
  \[ P(\theta) \propto 1 \quad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) \] 
  \[ \rightarrow \text{reduces to MLE objective} \]
Bayesian Learning for Thumbtacks

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

Likelihood function is Binomial:

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

• 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 – $P(\theta)$

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

- 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:** $\text{Beta}(\beta_H, \beta_T)$
- **Data:** $\alpha_H$ heads and $\alpha_T$ tails
- **Posterior distribution:**
  \[
P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)
  \]
Bayesian Posterior Inference

- Posterior distribution:

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

- Bayesian inference:
  - No longer single parameter
  - For any specific \( f \), the function of interest
  - Compute the expected value of \( f \)

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

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

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

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

- As more data is observed, Beta is more certain

- **MAP**: use most likely parameter to approximate the expectation

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

\[ E[f(\theta)] \approx f(\hat{\theta}) \]
MAP for Beta distribution

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

- MAP: use most likely parameter:

\[ \hat{\theta} = \arg \max_{\theta} P(\theta | D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to 1 \), 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?

\[
\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\mu} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right] = - \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0
\]

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

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

- Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(D \mid \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^3} \quad = 0
\]

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

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