Point Estimation

CSE 446: Machine Learning
Emily Fox
University of Washington
January 6, 2017

Maximum likelihood estimation
for a binomial distribution
Your first consulting job

• A bored Seattle billionaire 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:

  – You say: The probability is:
  – He says: Why???
  – You say: Because…

Thumbtack – Binomial distribution

• P(Heads) = \(\theta\), P(Tails) = 1-\(\theta\)
• Flips are i.i.d.:
  – Independent events
  – Identically distributed according to a binomial distribution

• Sequence D of \(\alpha_H\) heads (H) and \(\alpha_T\) tails (T)
• P(D | \(\theta\)) = 
The learning task

- Want to learn a model of thumbtack flips from experience

- **Example 1**: Maximum likelihood estimation
  What value of $\theta$ maximizes the likelihood of having seen the observed sequence (according to my model)?

- What is a likelihood function?

Maximum likelihood estimation

- **Data**: Observed set $D$ of $\alpha_H$ heads (H) and $\alpha_T$ tails (T)
- **Hypothesis**: Binomial distribution
- Learning $\theta$ is an optimization problem
  - What’s the objective function?

- **MLE**: Choose $\theta$ that maximizes the likelihood of observed data
  \[
  \hat{\theta} = \arg \max_{\theta} P(D \mid \theta) \\
  = \arg \max_{\theta} \ln P(D \mid \theta)
  \]
Your first 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: \( \frac{d}{d\theta} \ln P(D \mid \theta) = 0 \)

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: \( \theta = \frac{3}{5} \), 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???
• He says: Is this why I am paying you the big bucks???
Simple bound
(based on 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 \( \varepsilon > 0 \):
  \[
P(\left| \hat{\theta}_{MLE} - \theta^* \right| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}
  \]

PAC learning

- **PAC**: Probably Approximate Correct
- **Billionaire says**: I want to know the thumbtack parameter \( \theta \) within \( \varepsilon = 0.1 \), with probability at least \( 1 - \delta = 0.95 \). How many flips do I need?
  \[
P(\left| \hat{\theta}_{MLE} - \theta^* \right| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}
  \]
What about continuous-valued data?

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...
1D Gaussians

Fully specified by mean $\mu$ and variance $\sigma^2$ (or standard deviation $\sigma$)

Random variable the distribution is over e.g., height

1D Gaussian probability density function

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

Random variable the distribution is over e.g., height
Some properties of Gaussians

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

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

Learning a Gaussian

• Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., heights of students in class

• Learn parameters
  - Mean
  - Variance

\[
p(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
MLE for Gaussian

• Prob. of i.i.d. samples $D = \{x_1, \ldots, x_N\}$:

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

• 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 the mean?

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

- Again, set derivative to zero:

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

Learning Gaussian parameters

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

- FYI, MLE for the variance of a Gaussian is biased
  - Expected value of estimator is not true parameter!
  - Unbiased variance estimator:
    \[ \hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_{MLE})^2 \]
Recap of concepts

What you need to know...

• Learning is...
  - Collect some data
    • E.g., thumbtack flips
  - Choose a hypothesis class or model
    • E.g., binomial
  - Choose a loss function
    • E.g., data likelihood
  - Choose an optimization procedure
    • E.g., set derivative to zero to obtain MLE
  - Collect the big bucks

• Like everything in life, there is a lot more to learn...
  - Many more facets... Many more nuances...
  - The fun will continue...