Point Estimation

Zhang Hongxin
zhx@cad.zju.edu.cn

State Key Lab of CAD&CG, ZJU
2007-03-01
What you need to know

- Point estimation:
  - **Maximal Likelihood Estimation (MLE)**
  - Bayesian learning
  - **Maximize A Posterior (MAP)**
- Gaussian estimation
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off
Your first consulting job

- A billionaire from Beijing asks you a question:
  - B: I have thumbtack, if I flip it, what’s the probability it will fall with the nail up?
  - Y: Please flip it a few times …
  - Y: The probability is 3/5
  - B: Why???
  - Y: Because…
Binomial Distribution

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

$$P(D \mid \theta) = (1 - \theta)\theta\theta(1 - \theta)(1 - \theta)$$

- 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?
    \[
    D = \{T, H, H, T, T\}
    \]
- **MLE:** Choose $\hat{\theta}$ that maximizes the probability of observed data:
  \[
  \hat{\theta} = \arg \max_\theta P(D | \theta) \quad = \quad \arg \max_\theta \ln P(D | \theta) = ...
  \]
Maximum Likelihood Estimation (cont.)

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

Set derivative to zero:

\[ \frac{d}{d\theta} \ln P(D \mid \theta) = 0 \]
\[ \hat{\theta} = \frac{\alpha_T}{\alpha_H + \alpha_T} = \frac{3}{2 + 3} \]
How many flips do I need?

\[ \hat{\theta} = \frac{\alpha_T}{\alpha_H + \alpha_T} \]

- B: I flipped 2 heads and 3 tails.
- Y: \( \theta = 3/5 \), I can prove it!
- B: What if I flipped 20 heads and 30 tails?
- Y: Same answer, I can prove it!
- B: What’s better?
- Y: Humm… The more the merrier???
- B: Is this why I am paying you the big bucks???
Simple bound
(based on Höffding’s inequality)

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


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

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

$$N \geq \frac{1}{2\varepsilon^2} [\ln 2 - \ln \delta]$$

$$N \geq 270; (\varepsilon = 0.1, \delta = 0.01)$$
PAC Learning

- PAC: Probably Approximately Correct
- B: I want to know the thumbtack parameter $\theta$, within $\varepsilon = 0.1$, with probability at least $1-\delta = 0.99$. How many flips?
- Y: 270, 😊
Prior: knowledge before experiments

- B: Wait, I know that the thumbtack is “close” to 50-50. What can you …?
- Y: I can learn it the Bayesian way…

- Rather than estimating a single \( \theta \), we obtain a distribution over possible values of \( \theta \)

\[
D = \{T, H, H, T, T\}
\]

\[
P(\theta | D)
\]
Bayesian Learning

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

- Or equivalently:
  \[ P(\theta | D) \propto P(\theta)P(D | \theta) \]

Posterior  \[\downarrow\quad\text{Prior} \quad\downarrow\quad\text{Likelihood} \quad\downarrow\quad\text{Data distribution} \]
  \[\quad\rightarrow P(D) \quad\text{(Normalization constant)} \]
Bayesian Learning in our case

- 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 – \( P(\theta) \)

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

- Likelihood: Binomial distribution
  \[
P(D | \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}
  \]

- Posterior:
  \[
P(\theta | D) \propto P(\theta)P(D | \theta)
  \propto \theta^{\alpha_H} (1-\theta)^{\alpha_T} \theta^{\beta_H-1} (1-\theta)^{\beta_T-1}
  \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)
  \]
Using Bayesian posterior

- Posterior distribution:
  \[ P(\theta | D) \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T) \]

- Bayesian inference:
  - No longer single parameter:
    \[ E[f(\theta)] \sim \int_0^1 f(\theta)P(\theta | D)d\theta \]
  - Integral, 😞
MAP:
Maximum a posteriori approximation

\[ P(\theta | D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T) \]

\[ E[f(\theta)] = \int_{0}^{1} f(\theta) P(\theta | D) d\theta \]

- MAP: use most likely parameter

\[ \hat{\theta} = \arg \max_{\theta} P(\theta | D) \quad E[f(\theta)] \approx f(\hat{\theta}) \]
MAP for Beta distribution

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

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

- Beta prior equivalent to extra thumbtack flips
- As \( N = \alpha_T + \alpha_H \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
Gaussian distribution

Continuous variable:

\[ P(x \mid \mu, \delta) \sim \frac{1}{\delta \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}} \]

Consider the difference between continuous and discrete variables?
MLE for Gaussian

- Prob. of i.i.d. samples $D = \{x_1, x_2, \ldots, x_N\}$

likelihood

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

- The magic of log (to likelihood)

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

$$= -N \ln(\sigma \sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$
MLE for mean of a Gaussian

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

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

\[
\mu = \frac{1}{N} \sum_{i} x_i
\]
MLE for variance of a Gaussian

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

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

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

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

- **MLE**
  \[
  \hat{\mu} = \frac{1}{N} \sum x_i \\
  \hat{\sigma}^2 = \frac{1}{N} \sum (x_i - \mu)^2
  \]

- **Bayesian learning: prior?**

- **Conjugate priors:**
  - Mean: Gaussian priors
  - Variance: Wishart Distribution
Prediction of continuous variable

- B: Wait, that's not what I meant!
- Y: Chill out, dude.
- B: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- Y: I can regress that…

![Diagram showing a scatter plot with a linear regression line, correlating GPA with Salary.]
The regression problem

- **Instances:** \( <x_i, t_i> \)
- **Learn:** mapping from \( x \) to \( t(x) \).
- **Hypothesis space:** \( t(x) \approx \hat{f}(x) = \sum_{i=1}^{k} w_i h_i \)
  - Given, basis functions \( H = \{h_1, \ldots, h_k\} \)
  - Find coefficients \( w = \{w_1, \ldots, w_k\} \)
- **Problem formulation:**
  \[
  w^* = \arg \min_w \sum_j \left[ t(x_j) - \sum_{i=1}^{k} w_i h_i(x) \right]^2
  \]
But, why sum squared error?

- Model:

\[
P(t \mid x, w, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{[t - \sum_i w_i h_i(x)]^2}{2\sigma^2}}
\]

- Learn \( w \) using MLE
Maximizing log-likelihood

\[\ln P(D \mid w, \sigma) = \ln \prod_j \left( \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}} \right)\]

\[\Rightarrow \text{min} \sum_j \frac{[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}\]
Bias-Variance Tradeoff

- Choice of hypothesis basis introduce learning bias:
  - More complex basis:
    - Less bias
    - More variance (over-fitting)
What you need to know

- Point estimation:
  - Maximal Likelihood Estimation
  - Bayesian learning
  - Maximal a Posterior

- Gaussian estimation

- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians

- Bias-Variance trade-off
Homework

- Finish the “Gaussian parameters learning”
  - Please use google, ^_^*