Point Estimation and Maximum Likelihood

CSE512 – Machine Learning, Spring 2018, Stony Brook University

Instructor: Minh Hoai Nguyen (minhhoai@cs.stonybrook.edu)

Date: 24 Jan 2018

Many slides are by Carlos Guestrin at University of Washington

Your First Consulting Job

- The company of a billionaire from Long Island is hiring. He receives too many applications that he cannot review them all. He consults you.
- You say: just flip a coin.

- He says: Not enough. A coin is 50/50 chance, so I still have to review 50% of the applications
- You say: flip a thumbtack instead.
- He says: good idea. I have one right here.

Your First Consulting Job

- He says: wait, if I flip it, what's the probability it will fall with the head 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) = θ, P(Tails) = 1-θ

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your first learning algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ \ln P(\mathcal{D} \mid \theta) = 0}$

How many flips do I need?

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 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: Hmm… 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 θ^* be the true parameter, for any $\varepsilon > 0$:

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

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter θ, within $\varepsilon = 0.1$, with probability at least $1-\delta = 0.95$. How many flips?

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

What about continuous variables?

- Billionaire says: If I want to predict the amount of fuel to fly from New York to Chicago, what can you do for me? It's a continuous variable!
- You say: Let me tell you about Gaussians...

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

Gaussian Distribution Reviewed

Probability density function (pdf): $P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Notation: $x \sim \mathcal{N}(\mu, \sigma^2)$

Some Properties of Gaussians

Affine transformation

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$$y = ax + b \Rightarrow y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

Sum of Gaussians

$$y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$x \sim \mathcal{N}(\mu_x, \sigma_x^2)$$

$$z = x + y \Rightarrow z \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_x^2)$$

Learning a Gaussian

- Given data samples $D = \{x_1, \cdots, x_n\}$

- Suppose the data comes from a Gaussian $\mathcal{N}(\mu, \sigma^2)$
- Learn the parameters of the Gaussian

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

Maximum Likelihood Estimator (MLE)

The likelihood of data (probability)

$$P(D|\mu, \sigma) =$$

The log-likelihood

$$\log(P(D|\mu, \sigma)) =$$

MLE for Mean

$$\frac{\partial \log(P(D|\mu, \sigma))}{\partial \mu} =$$

MLE for Variance

$$\frac{\partial \log(P(D|\mu,\sigma))}{\partial \sigma} =$$

MLE for Gaussian parameters

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Biased and Unbiased Estimators

MLE of the mean is un-biased:

$$E_D[\hat{\mu}_{MLE}] = \mu$$

Expectation over all dataset *D* of n elements

Proof:

$$\begin{aligned} E_D[\hat{\mu}_{MLE}] &= E_D[\frac{1}{n}\sum_{i=1}^{n} x_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\ &= \mu \end{aligned}$$

Biased and Unbiased Estimators

MLE of the variance is biased:

$$E_D[\hat{\sigma}_{MLE}] \neq \sigma$$

Expectation over all dataset *D* of n elements

The unbiased estimator of variance

$$\hat{\sigma}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Proof: exercise!

Maximum A Posterior (MAP)

MLE:

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \, P(D|\mu, \sigma)$$

Frequentist

MAP:

Bayesian

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \, P(\mu, \sigma|D)$$

MAP

$$P(\mu,\sigma|D) = \frac{P(D|\mu,\sigma)P(\mu,\sigma)}{P(D)}$$

$$\propto P(D|\mu,\sigma)P(\mu,\sigma)$$

MLE = MAP if we assume uniform prior for (μ, σ)

MAP for Gaussian

Use conjugate priors for the parameters

- Mean: Gaussian prior
- Variance: Wishart Distribution

Prior for Mean

$$P(\mu) = \mathcal{N}(\mu|\mu_0, \lambda^2)$$

MAP for Mean of Gaussian

$$\log(P(D|\mu, \sigma)P(\mu)) = \log(P(D|\mu, \sigma)) + \log(P(\mu))$$

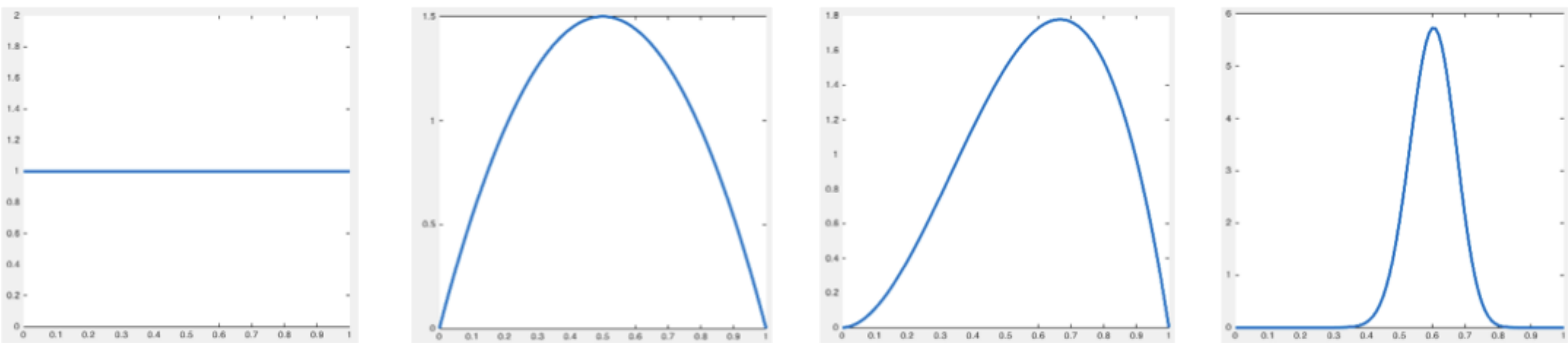
To find MAP estimate, take derivative and set it to 0

Prior for Thumbtack Problem

Data likelihood: $$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If we use Beta prior for the parameter $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$$

Beta(1,1) Beta(2,2) Beta(3,2) Beta(30,20)



MAP estimate for Thumbtack Problem

Data likelihood: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

If we use Beta prior for the parameter $P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$

What is Machine Learning? – Revisited

- Machine Learning is…
  - Collect some data
    - E.g., thumbtack flips, fuel consumption
  - Choose a hypothesis class or model
    - E.g., binomial, Gaussian
  - Choose a loss function
    - E.g., data likelihood, parameter likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE

Lecture Summary

- What Machine Learning is
- Maximum Likelihood Estimator (MLE)
- Maximum A Posterior Estimator (MAP)
- Binomial Distribution
- Gaussian Distribution