MLE & MAP

INSTRUCTOR: HONGJIE CHEN

MAY 26TH 2022

Deal with Data Sparsity

- Estimate probabilities from sparse data
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriori estimation (MAP)

- Represent joint probability distribution other than using tables
  - Graphical models

A Toy Example to Explain MLE & MAP

- Given a coin, estimate the probability that it turns up heads $(X = 1)$ or tails $(X = 0)$
- Test A: 3 flips, 2 heads $(X = 1)$, 1tail $(X = 0)$
- Test B: 100 flips, 51 heads $(X = 1)$, 49 tails $(X = 0)$
- Test C: Keep flipping and develop an online learning algorithm that gives reasonable estimate for each flip

Estimating Probabilities

- Maximum Likelihood Estimation (**MLE**)
  - Choose parameter θ that maximizes $P(\text{data} \mid \theta)$
  - $$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$
- Maximum A Posteriori estimation (**MAP**)
  - Choose parameter θ that maximizes $P(\theta \mid \text{data})$
  - $$\hat{\theta} = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

Confusing?
Prior: $P(\theta)$

MLE - $P(data \mid \theta)$

- α_1: # head up, α_0: # tail up
- $P(X = 1) = \theta, P(X = 0) = 1 - \theta$
- Data D: 1, 0, 0, 1, 1
- $P(D \mid \theta)$: $\theta, 1 - \theta, 1 - \theta, \theta, \theta = \theta^{\alpha_1}(1 - \theta)^{\alpha_0}$ **i.i.d.~Bernoulli**

- Our goal: $\hat{\theta} = \text{argmax}_\theta P(D \mid \theta)$
- How to obtain the extreme value for a function?

Derivative

- Our goal: $\hat{\theta} = \text{argmax}_{\theta} P(D \mid \theta)$
- $P(D \mid \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$
- Calculate θ from $\frac{d}{d\theta} P(D \mid \theta) = 0$ (Show me the maths!)
  - An easier route with $lnP(D \mid \theta)$
- $\hat{\theta}_{\text{MLE}} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$ **Questions: How to get a more accurate $\hat{\theta}_{\text{MLE}}$? Think about # flips**

Prepare for MAP

- Recall Bayes theorem
  - $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$
  - $$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)P(\theta)}{P(\text{data})}$$

MAP

- Bayes theorem

$$\underbrace{P(\theta \mid \text{data})}_{\text{Posteriori}} = \frac{\overbrace{P(\text{data} \mid \theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\text{data})}$$

No variable in data

- Equivalently, $P(\theta \mid \text{data}) \propto P(\text{data} \mid \theta)P(\theta)$

Prior Probability: $P(\theta)$

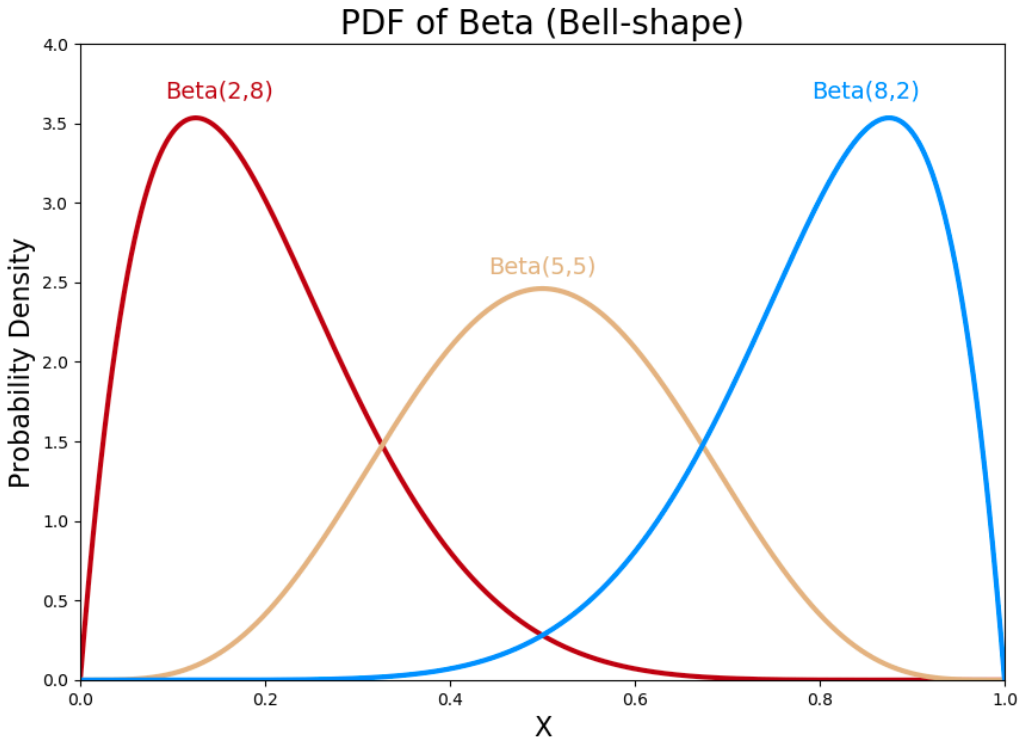
- Let's review what θ and $P(\theta)$ mean
  - θ: the probablity a coin turns head up in one flipping
  - $P(\theta)$: the distribution of this probability parameter

Beta Distribution

- We choose Beta distribution as the prior

$$P(\theta) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_1, \beta_0)} \sim Beta(\beta_1, \beta_0)$$

A normalization constant



Why Beta Distribution as Prior

- Reasonable, and easy to compute

$$P(\theta) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_1, \beta_0)}, \qquad P(D \mid \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$P(\theta \mid \mathrm{D}) \propto P(\mathrm{D} \mid \theta) P(\theta)$$

$$= \frac{\theta^{\alpha_1 + \beta_1 - 1}(1-\theta)^{\alpha_0 + \beta_0 - 1}}{B(\alpha_1 + \beta_1, \alpha_0 + \beta_0)} \Rightarrow \hat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_\theta P(\theta \mid D) = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

As $N \to \infty$, prior is dominated when $N = (\alpha_0 + \alpha_1) \gg \beta_0 + \beta_1$

Questions about Prior

- Are other prior probability distributions possible?

- How to set parameters for the prior, e.g., for Beta distribution?

Thoughts

- Are other prior probability distributions possible?

  Yes. substitute $P(\theta)$ and calcualte derivative to get $\hat{\theta}_{\mathbf{MAP}}$

- How to set parameters for the prior, e.g., for Beta distribution?

  Reparametrize β_1, β_0 to other parameters such as mean and variance μ, σ^2, conduct pilot experiments to obtain μ and σ^2

$$\nu := \beta_1 + \beta_0 = \frac{\mu(1-\mu)}{\sigma^2} - 1, \begin{cases} \beta_1 = \mu\nu \\ \beta_0 = (1-\mu)\nu \end{cases}$$

New parameters to estimate :(

Conjugate Prior

- If the posterior distribution $P(\theta|D)$ and the prior distribution $P(\theta)$ are in the same probability distribution family.

Likelihood: $P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0} \sim Binomial(\alpha_1, \alpha_0)$

Prior: $P(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_1, \beta_0)} \sim Beta(\beta_1, \beta_0)$

Posterior: $P(\theta|D) = \frac{\theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}}{B(\alpha_1+\beta_1, \alpha_0+\beta_0)} \sim Beta(\alpha_1+\beta_1, \alpha_0+\beta_0)$

For Binomial likelihood, conjugate prior is Beta distribution

Conjugate Prior for Multinomial

- Let's use a dice instead of a coin (k outcomes instead of 2)

  Likelihood: $P(D \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\cdots\theta_k^{\alpha_k} \sim Multinomial(\theta_1, \theta_2, \ldots, \theta_k)$

  If the prior is a Dirichlet distribution: $P(\theta) = \frac{\prod_{i=1}^{k}\theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim Dirichlet(\beta_1, \ldots, \beta_k)$

  Then the posterior is a Dirichlet distribution:

  $$P(\theta \mid D) \sim Dirichlet(\alpha_1 + \beta_1, \ldots, \alpha_k + \beta_k)$$

  For Multinomial likelihood, conjugate prior is Dirichlet distribution