CS 581 – Advanced Artificial Intelligence

Topic: Parameter Estimation

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

Parameter Estimation

- Imagine we have a thumbtack
- Flip it, and it comes as heads or tails

- P(Heads) = θ, P(Tails) = 1- θ
- Assume we flip it $(a + b)$ times and it comes head a times. What is θ if
  - $a = 4, b = 6$
  - $a = 42, b = 58$
  - $a = 407\ b = 593$
- Can you prove your answers?
- Can you associate a confidence score with your estimates?

We will see two approaches

1. Maximum likelihood estimation
2. Bayesian estimation

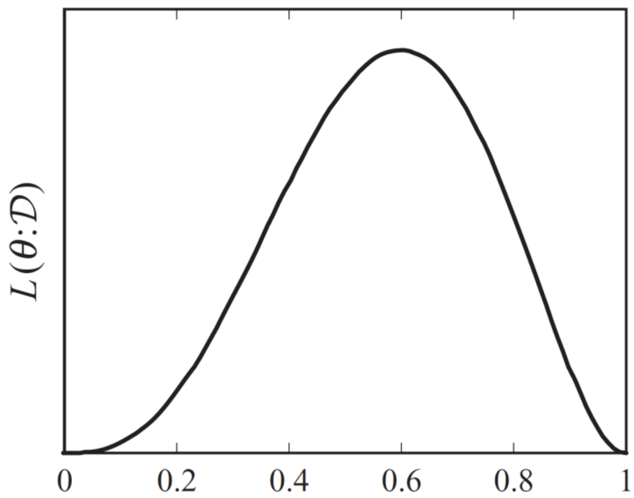
Maximum Likelihood Estimation

MAXIMUM LIKELIHOOD ESTIMATION (MLE) – FOR THE THUMBTACK EXAMPLE

- Assume we have a set of thumbtack tosses
  - $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$ where a are Heads and b are Tails
- Hypothesis space: $[0, 1]$
- Find a "good" scoring $\theta \in [0, 1]$
- Let $f(\theta : \mathcal{D})$ be the score of θ given $\mathcal{D}$, where a high score is a "good" score
- Learning: $\underset{\theta \in [0,1]}{\operatorname{argmax}} f(\theta : \mathcal{D})$
- What is $f(\theta : \mathcal{D})$?
- The space [0, 1] is infinite. How do we search this space efficiently?

LIKELIHOOD

- What is the probability, or *likelihood*, of seeing the sequence H, T, T, H, H?
  - $\theta * (1 - \theta) * (1 - \theta) * \theta * \theta = \theta^3 (1 - \theta)^2$



When is $L(\theta : \mathcal{D})$ maximum?

Likelihood/log-likelihood

- Number of heads = a, number of tails = b
- Likelihood: $L(\theta:\mathcal{D}) = \theta^a(1-\theta)^b$
- Log-likelihood: $l(\theta:\mathcal{D}) = a\log\theta + b\log(1-\theta)$
- Note that $L(\theta:\mathcal{D})$ achieves its maximum for θ that maximizes $l(\theta:\mathcal{D})$
- Find θ that maximizes the log-likelihood
- Take derivate of $l(\theta:\mathcal{D})$ w.r.t. θ and set it to zero

Bayesian Estimation

Bayesian Estimation

- MLE gives the same estimate of $\theta = 0.4$, if we have 4 H and 6 T, as well as 4M H and 6M T
- In Bayesian estimation, rather than a single θ,
  - We assume a prior belief about θ: $\boldsymbol{p(\theta)}$, and we estimate
    - The posterior distribution over θ: $\boldsymbol{p(\theta \mid \mathcal{D})}$
    - The probability distribution for the next toss: $\boldsymbol{P(d_{m+1} \mid \mathcal{D})}$

POSTERIOR: $p(\theta \mid \mathcal{D})$

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta)P(\mathcal{D} \mid \theta)}{P(\mathcal{D})}$$

$P(\mathcal{D})$ does not depend on θ. Hence, it can be treated as a constant from the perspective of θ.

$$p(\theta \mid \mathcal{D}) \propto p(\theta)P(\mathcal{D} \mid \theta)$$

Next, assume each data point is independent given θ: $d_i \perp d_j \mid \theta$

$$P(\mathcal{D}|\theta) = P(d_1|\theta)P(d_2|\theta)P(d_3|\theta)\cdots P(d_m|\theta) = \prod_{i=1}^{m} P(d_i \mid \theta)$$

Hence, the posterior becomes

$$p(\theta \mid \mathcal{D}) \propto p(\theta)\prod_{i=1}^{m} P(d_i \mid \theta)$$

PREDICTION: $P(d_{m+1} \mid \mathcal{D})$

$$P(d_{m+1}|D) = \int_0^1 P(d_{m+1}|\theta, \mathcal{D})p(\theta|\mathcal{D})d\theta$$

Assuming $d_i \perp d_j \mid \theta$:

$$P(d_{m+1}|D) = \int_0^1 P(d_{m+1}|\theta)p(\theta|\mathcal{D})d\theta$$

Using the posterior equation from the previous slide:

$$P(d_{m+1}|D) \propto \int_0^1 P(d_{m+1}|\theta)p(\theta)\prod_{i=1}^{m} P(d_i \mid \theta)\, d\theta$$

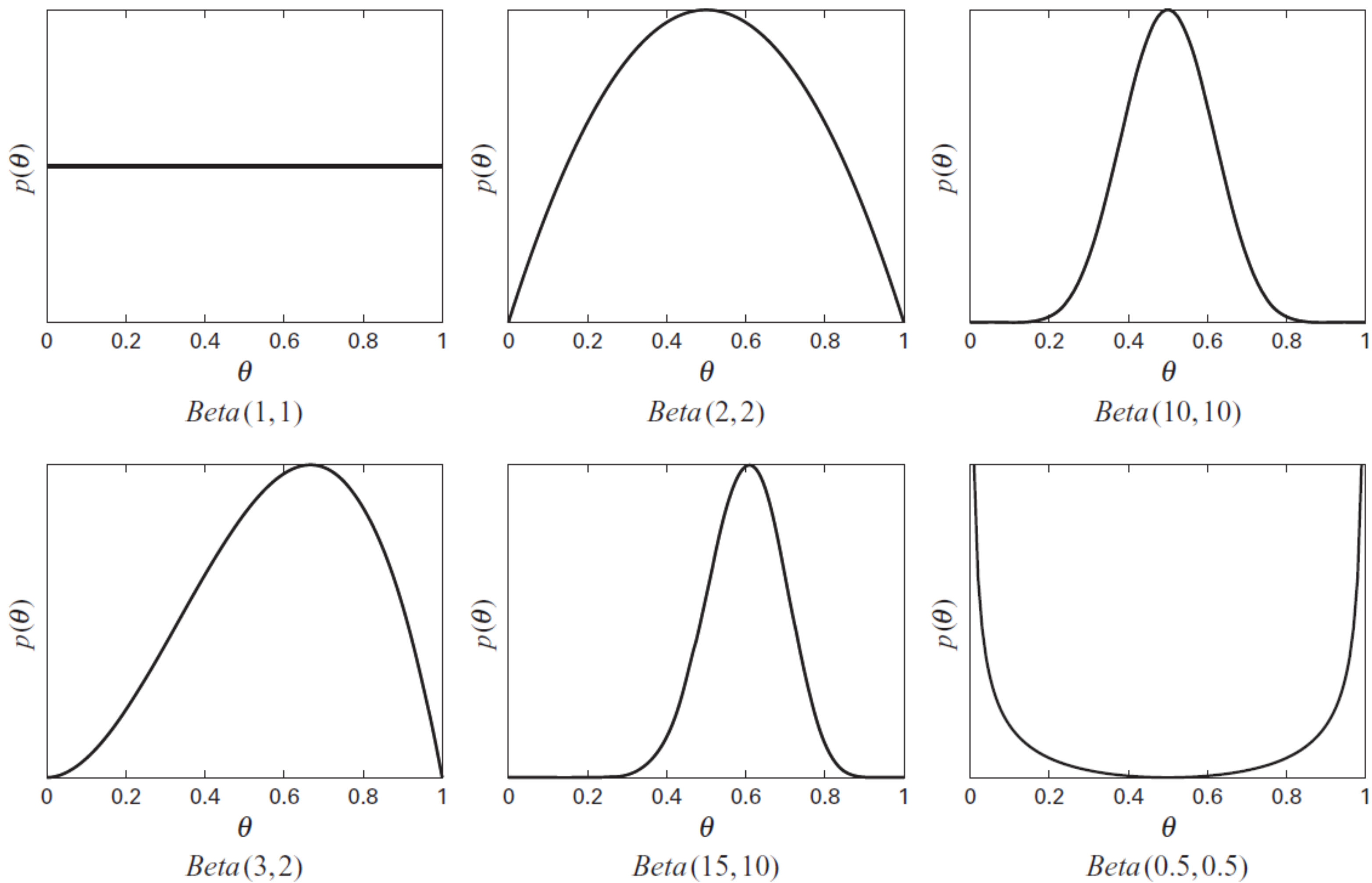
UNIFORM PRIOR

- Assume a Heads and b Tails
- Assume a uniform prior over θ. That is, $p(\theta) = 1$
- What is $P(d_{m+1} = Heads \mid \mathcal{D})$?
  - (a+1)/(a+b+2)
- What is $p(\theta \mid \mathcal{D})$?
  - Beta(a+1, b+1)

Beta Distribution

- $\theta \sim \mathrm{Beta}(\alpha,\beta)$ if $p(\theta) = \gamma\theta^{\alpha-1}(1-\theta)^{\beta-1}$ where γ is a normalizing constant
- Mean: $\alpha/(\alpha+\beta)$
- Mode: $(\alpha-1)/(\alpha+\beta-2)$
- Note that the mode is closer to the mean when α and β are large
- Read more at
  - https://en.wikipedia.org/wiki/Beta_distribution

BETA DISTRIBUTION



BETA DISTRIBUTION

- What is $P(d_{M+1}=True \mid d_1,\ldots,d_M)$ if the prior is Beta(α,β)?
  - $P(X[M+1]=\text{True} \mid D) = (a + \alpha) / (a + b + \alpha + \beta)$
- What is the posterior, $P(\theta \mid D)$, if the prior is Beta(α,β)?
  - $P(\theta \mid D) = \text{Beta}(a + \alpha,\ b + \beta)$
- α and β work like pseudo-counts for the positive and negative cases respectively
- What values to choose for α and β?
  - It depends on our belief and the strength of our belief

Dirichlet priors

- Generalizes the Beta distribution for multinomials

$$\theta \sim Dirichlet(\alpha_1, \dots, \alpha_K) \text{ if } P(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- What is $P(d_{M+1}=v_i \mid D)$ if the prior is Dirichlet?
  - $P(d_{M+1}=v_i \mid D) = (n_i+\alpha_i) / (|D| + \alpha)$ where n_i is the number of times the i^{th} case appears in D and $\alpha = \alpha_1 + \alpha_2 + \dots + \alpha_K$
- What is the posterior, $P(\theta \mid D)$, if the prior is Dirichlet?
  - $P(\theta \mid D) = Dirichlet(n_1 + \alpha_1, n_2 + \alpha_2, \dots, n_K + \alpha_K)$