Probability & Statistics for DS & AI

Maximum a posteriori estimation

Michele Guindani

Summer

Adding prior information into the estimation problem

- In ML estimation, the parameter $\boldsymbol{\theta}$ is treated as a deterministic quantity.
- There are, however, many situations where we have some prior knowledge about $\boldsymbol{\theta}$.

Adding prior information into the estimation problem

- In ML estimation, the parameter $\boldsymbol{\theta}$ is treated as a deterministic quantity.
- There are, however, many situations where we have some prior knowledge about $\boldsymbol{\theta}$.
- For example, we may not know exactly the speed of a car, but we may know that the speed in Mumbai is roughly 35 km/h.
- As another example, we may know that in his career, LeBron James is a 34.5 percent 3-point shooter.

Adding prior information into the estimation problem

- In ML estimation, the parameter $\boldsymbol{\theta}$ is treated as a deterministic quantity.
- There are, however, many situations where we have some prior knowledge about $\boldsymbol{\theta}$.
- For example, we may not know exactly the speed of a car, but we may know that the speed in Mumbai is roughly 35 km/h.
- As another example, we may know that in his career, LeBron James is a 34.5 percent 3-point shooter.

⇨ We may want to incorporate this prior knowledge into the problem (Bayesian)

Adding prior information into the estimation problem

- In ML estimation, the parameter $\boldsymbol{\theta}$ is treated as a deterministic quantity.
- There are, however, many situations where we have some prior knowledge about $\boldsymbol{\theta}$.
- For example, we may not know exactly the speed of a car, but we may know that the speed in Mumbai is roughly 35 km/h.
- As another example, we may know that in his career, LeBron James is a 34.5 percent 3-point shooter.

⇨ We may want to incorporate this prior knowledge into the problem (Bayesian)

- We can do this by considering a prior distribution on $\boldsymbol{\theta}$.

Adding prior information into the estimation problem

- In ML estimation, the parameter $\boldsymbol{\theta}$ is treated as a deterministic quantity.
- There are, however, many situations where we have some prior knowledge about $\boldsymbol{\theta}$.
- For example, we may not know exactly the speed of a car, but we may know that the speed in Mumbai is roughly 35 km/h.
- As another example, we may know that in his career, LeBron James is a 34.5 percent 3-point shooter.

⇨ We may want to incorporate this prior knowledge into the problem (Bayesian)

- We can do this by considering a prior distribution on $\boldsymbol{\theta}$.

  For example,

$$\boldsymbol{\theta} = \text{average speed of cars in Mumbai} \sim N(35, 5^2) \quad \text{[Expert Assessment]}$$

Maximum a posteriori (MAP) estimate

The idea behind the MAP estimate is to find that value of θ that maximizes the information on the parameter of interest based on the prior information and the data:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} \operatorname*{argmax}_{\boldsymbol{\theta}} = \left\{ \underbrace{\log f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta})}_{\text{information from the data}} + \underbrace{\log f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}_{\text{information from the prior}} \right\}$$

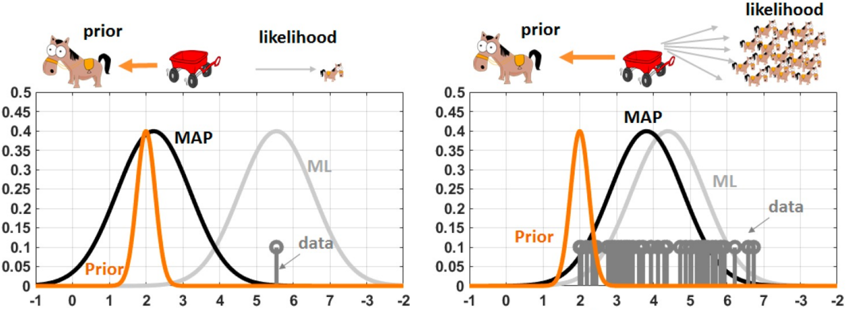
Maximum a posteriori (MAP) estimate

The idea behind the MAP estimate is to find that value of θ that maximizes the information on the parameter of interest based on the prior information and the data:

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} \underset{\boldsymbol{\theta}}{\operatorname{argmax}} = \left\{ \underbrace{\log f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta})}_{\text{information from the data}} + \underbrace{\log f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}_{\text{information from the prior}} \right\}$$

- The prior works as a regularization of the estimation problem, or penalization of the likelihood
- These types of penalizations are common in ML

Sample size and Inference

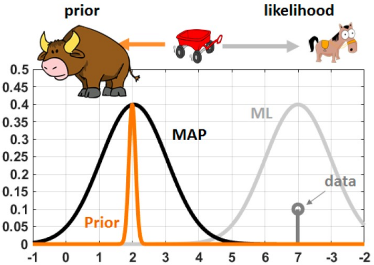


(a) $N = 1$ (b) $N = 50$

The subfigures show the prior distribution $f_\Theta(\theta)$ and the likelihood function $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$, given the observed data. (a) When $N = 1$, the estimated posterior distribution $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ is pulled towards the prior. (b) When $N = 50$, the posterior is pulled towards the ML estimate. The analogy for the situation is that each data point is acting as a small force against the big force of the prior. As N grows, the small forces of the data points accumulate and eventually dominate.

Prior strength and Inference

The inference is also affected by the strength of the prior information. Let σ_0 indicate the prior variance



(a) $\sigma_0 = 0.1$ (b) $\sigma_0 = 1$

The subfigures show the prior distribution $f_\Theta(\theta)$ and the likelihood function $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$, given the observed data. (a) When $\sigma_0 = 0.1$, the estimated posterior distribution $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ is pulled towards the prior. (b) When $\sigma_0 = 1$, the posterior is pulled towards the ML estimate. An analogy for the situation is that the strength of the prior depends on the magnitude of σ_0. If σ_0 is small the prior is strong, and so the influence is large. If σ_0 is large the prior is weak, and so the ML estimate will dominate.