

Introduction to Statistical Learning

Omid Safarzadeh

January 19, 2022

Table of contents

- 1 Likelihood and posterior distribution
 - Computing the posterior
 - Maximum likelihood estimation (MLE)
- 2 Maximum a posteriori (MAP) estimation
 - Posterior mean
 - MAP properties
- 3 Bayesian linear regression
- 4 Reference

Likelihood and posterior distribution

- $X_i \sim \text{Ber}(\theta)$
- Θ : Probability of heads (uncertain value).
- N : Number of coin flips
- $\mathcal{D} := \{N_1 \text{ heads}, N_0 \text{ tails}\}$ (observed data)
- \mathbf{D} : Random variable that represents data, i.e., random number of heads and tails given N coin flips

Likelihood and posterior distribution

likelihood:

$$p_{\mathbf{D}|\Theta}(\mathcal{D}|\theta) = \binom{N_1 + N_0}{N_1} \theta^{N_1} (1 - \theta)^{N_0}$$

Posterior distribution:

$$\begin{aligned} p_{\Theta|\mathbf{D}}(\theta|\mathcal{D}) &= \frac{p_{\mathbf{D}|\Theta}(\mathcal{D}|\theta) p_{\Theta}(\theta)}{p_{\mathbf{D}}(\mathcal{D})} \\ &= \underbrace{\frac{\binom{N_1 + N_0}{N_1}}{p_{\mathbf{D}}(\mathcal{D})}}_{\text{constant}} \theta^{N_1} (1 - \theta)^{N_0} p_{\Theta}(\theta) \\ &\propto \theta^{N_1} (1 - \theta)^{N_0} \underbrace{p_{\Theta}(\theta)}_{\text{prior}} \end{aligned}$$

Computing the posterior

- We want a close-form expression for $p_{\Theta|\mathbf{D}}(\theta|\mathcal{D})$
- Take $\Theta \sim \text{Beta}(a, b)$
- Recall:

$$\Theta \sim \text{Beta}(a, b) \Rightarrow P_{\Theta}(\theta) = \underbrace{\frac{1}{B(a, b)}}_{\text{constant}} \theta^{a-1} (1 - \theta)^{b-1} \quad \text{for } \theta \in [0, 1]$$

Hence:

$$\begin{aligned} p_{\Theta|\mathbf{D}}(\theta|\mathcal{D}) &\propto \theta^{N_1} (1 - \theta)^{N_0} \theta^{a-1} (1 - \theta)^{b-1} \\ &\propto \theta^{N_1+a-1} (1 - \theta)^{N_0+b-1} \\ &\Rightarrow \Theta|\mathcal{D} \sim \text{Beta}(N_1 + a, N_0 + b) \end{aligned}$$

Computing the posterior

- $\text{Beta}(a,b)$: our prior belief about Θ
- Nothing known a priori: $a = 1, b = 1 \Rightarrow \text{Beta}(1,1) = \text{Unif}([0,1])$

Maximum likelihood estimation (MLE)

Goal: Infer Θ from \mathcal{D}

MLE:

$$p(\mathcal{D}|\theta) \propto \theta^{N_1}(1 - \theta)^{N_0}$$

$$\hat{\theta}_{MLE} := \arg_{\theta} \max p(\mathcal{D}|\theta)$$

For the example:

$$\hat{\theta}_{MLE} := \arg_{\theta \in [0,1]} \max \theta^{N_1}(1 - \theta)^{N_0}$$

$$\hat{\theta}_{MAP} := \arg_{\theta} \max p(\theta | \mathcal{D})$$

For the example:

$$\hat{\theta}_{MAP} = \arg_{\theta \in [0,1]} \max \theta^{a+N_1} (1-\theta)^{b+N_0-1}$$

What happens when we start with a uniform prior, i.e., Beta(1; 1)?

Posterior mean

Posterior mean: $E[\Theta|\mathcal{D}]$

For the example:

$$E[\Theta|\mathcal{D}] = \frac{a + N_1}{a + b + N_0 + N_1}$$

Since for $X \sim \text{beta}(x, y)$ we have

$$E[X] = \frac{x}{x + y}$$

Hence, in general $\hat{\theta}_{MLE}$, $\hat{\theta}_{MAP}$ and $E[\Theta|\mathcal{D}]$ are different.

Is MAP a good estimate?

- MAP = point estimate (does not measure uncertainty)

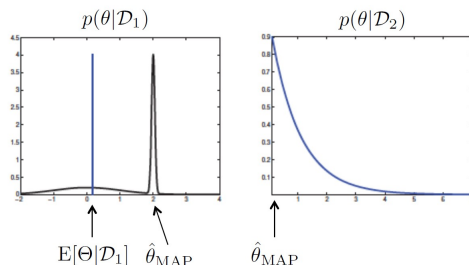


Figure: Murphy, 2012

What does MAP really optimize?

- Assume θ is the true parameter (realization Θ)
- Loss: $L(\theta, \hat{\theta}) = I(\theta \neq \hat{\theta})$
- $\hat{\theta}_{\text{MAP}}$ is the optimal estimate

Bayesian linear regression

- Least squares, ridge regression, lasso all produce point estimates, i.e., they output a single solution (least squares = MLE, ridge and lasso = posterior mode (MAP))
- Bayesian linear regression provides a posterior for β
- We can specify any prior and likelihood on β !
- But we assume that both prior and likelihood is Gaussian for analytical tractability!

Gaussian likelihood:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

What is the likelihood of $\mathbf{Y} = \mathbf{y}$ given that the true parameter vector is β and data \mathbf{X} is observed?

$$\begin{aligned} L(\beta) &= p(\mathcal{D}|\beta) \quad \underbrace{\quad}_{\text{Since } \mathbf{X} \text{ is fixed}} \quad p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1}^n p(y_i|\mathbf{x}_i, \beta) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2}\right) \\ &\sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \end{aligned}$$

- \mathbf{I}_n is the $n * n$ identity matrix

Gaussian posterior

- Gaussian prior + Gaussian likelihood \Rightarrow Gaussian posterior

General formula for Gaussian posterior:

- 1 $p(\beta) \sim \mathcal{N}(\mu_0, \Sigma_0)$
- 2 $p(\mathcal{D}|\beta) \sim \mathcal{N}(\mathbf{X}\beta, \Sigma_{\mathcal{D}})$
- 3 1&2 implies that $p(\beta|\mathcal{D}) \sim \mathcal{N}(\mu_{\beta|\mathcal{D}}, \Sigma_{\beta|\mathcal{D}})$

We have

- $\Sigma_{\beta|\mathcal{D}}^{-1} = \Sigma_0^{-1} + \mathbf{X}^T \Sigma_{\mathcal{D}}^{-1} \mathbf{X}$
- $\mu_{\beta|\mathcal{D}} = \Sigma_{\beta|\mathcal{D}} (\mathbf{X}^T \Sigma_{\mathcal{D}}^{-1} \mathbf{y} + \Sigma_0^{-1} \mu_0)$

Example 3.1

- How can we use $p(\beta|\mathcal{D}) \sim \mathcal{N}(\mu_{\beta|\mathcal{D}}, \Sigma_{\beta|\mathcal{D}})$?
- Assume we trained our model and fixed $p(\beta|\mathcal{D})$. We can use this to learn the distribution of Y given that we observe a new data instance \mathbf{x} .

Posterior predictive density at test point \mathbf{x} :

$$p(y|\mathbf{x}, \mathcal{D}) = \int_{\beta} \underbrace{p(y|\mathbf{x}, \beta)}_{\sim \mathcal{N}(\mathbf{x}^T \beta, \sigma^2)} p(\beta, \mathcal{D}) d\beta$$
$$\Rightarrow Y \sim \mathcal{N}(\underbrace{\mu_{\beta|\mathcal{D}}^T \mathbf{x}}_{\text{mean}}, \underbrace{\sigma^2 + \mathbf{x}^T \Sigma_{\beta|\mathcal{D}} \mathbf{x}}_{\text{variance}})$$

- σ^2 : variance of the noise term
- $\Sigma_{\beta|\mathcal{D}}$: covariance of the parameters

Example 3.1 cont.

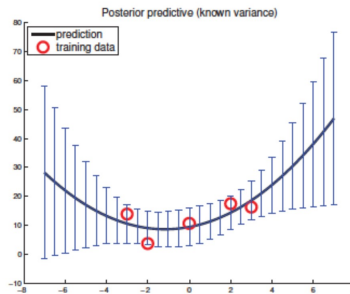


Figure: Murphy, 2012

- Y-axis: Y values, X-axis: X values
- Red circles are training points
- Black curve is the posterior mean of Y given \mathcal{D} and $X = x$
- Error bars (vertical bars): two standard deviations range for the posterior predictive density

Example 3.1 cont.

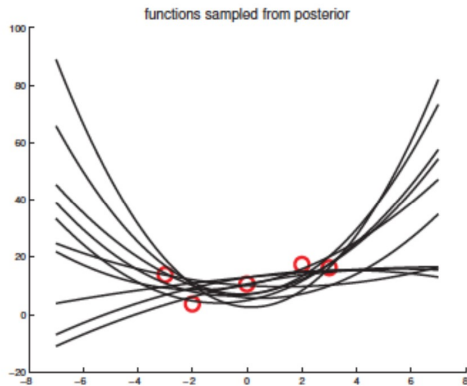


Figure: Murphy, 2012

Murphy, K. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
<https://books.google.fr/books?id=NZP6AQAAQBAJ>