Introduction to Statistical Learning

Omid Safarzadeh

January 18, 2022

Table of contents

Likelihood and posterior distribution

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

Likelihood and posterior distribution

likelihood:

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

Posterior distribution:

$$p_{\Theta|\mathbf{D}}(\theta|\mathscr{D}) = \frac{p_{\mathbf{D}|\Theta}(\mathscr{D}|\theta) p_{\Theta}(\theta)}{p_{\mathbf{D}}(\mathscr{D})}$$

$$= \underbrace{\frac{\binom{N_1+N_0}{N_1}}{p_{\mathbf{D}}(\mathscr{D})}}_{\text{constant}} \theta^{N_1} (1-\theta)^{N_0} p_{\Theta}(\theta)$$

$$\propto \theta^{N_1} (1-\theta)^{N_0} \underbrace{p_{\Theta}(\theta)}_{prior}$$

Computing the posterior

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

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

Hence:

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

Computing the posterior

- 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 $\mathscr{D}$
MLE:

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

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

For the example:

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

MAP estimation

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

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: $E[\Theta|\mathscr{D}]$
For the example:

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

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 $E[\Theta|\mathscr{D}]$ are different.

Is MAP a good estimate?

- MAP $=$ point estimate (does not measure uncertainty)

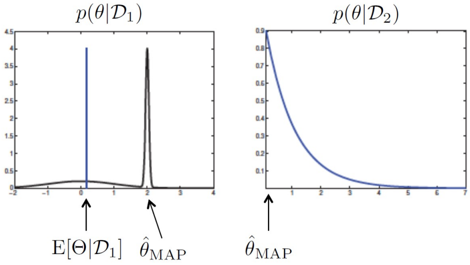


Figure: Murphy, 2012

What does MAP really optimize?

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

Ridge regression from a Bayesian perspective

Posterior distribution:

$$p(\beta, \mathscr{D}) \propto p(\mathscr{D}|\beta)p(\beta)$$

- Working with log is more convinient and numerically efficient

Ridge regression from a Bayesian perspective

Maximum aposteriori estimate of β:

$$\hat{\beta}^{MAP} = \arg_\beta \max p(\beta|\mathscr{D})$$

$$= \arg_\beta \max \left[\log p(\beta|\mathscr{D})\right]$$

$$= \arg_\beta \min \left[-\log p(\beta|\mathscr{D})\right]$$

Ridge regression from a Bayesian perspective

Solution:

$$-\log p(\beta|\mathscr{D}) \propto \frac{1}{2\sigma^2} \underbrace{(\sum_{i=1}^{n}(y_i - \beta^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2)}_{\text{Loss}_{\text{Ridge}}(\beta,\lambda)}$$

$$\underbrace{- n\log(\frac{1}{\sqrt{2\pi\sigma}}) - p\log(\frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma}})}_{\text{indep. of}\beta}$$

$$\Rightarrow \arg_\beta \min[-\log p(\beta|\mathscr{D})] = \arg_\beta \min Loss_{Ridge}(\beta, \lambda)$$

- Hence, $\hat{\beta}^{MAP} = \hat{\beta}^{Ridge}$ under prior $\beta \sim \mathscr{N}(0, \frac{\sigma^2}{\lambda}\mathbf{I}_p)$ and Gaussian likelihood
- $\mathbf{I}_p$ is the p by p identity matrix

Lasso from a Bayesian perspective

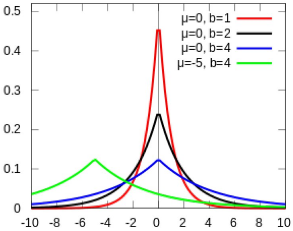
Prior: Laplace prior for β

$$\beta_j \sim Lap(0, \frac{2\sigma^2}{\lambda}) \Rightarrow p(\beta_j) = \frac{\lambda}{4\sigma^2} exp(-\frac{\lambda}{2\sigma^2}|\beta_j|)$$

β_j, $j = 1, ..., p$ are i.i.d.

- If $Z \sim Lap(\mu, b)$, then $E[Z] = \mu$, Var(Z)=$2b^2$,

$$p(Z) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b})$$



Likelihood: Gaussian likelihood

Lasso from a Bayesian perspective

Maximum aposteriori estimate of β:

$$\hat{\beta}^{MAP} = \text{arg}_{\beta} \max \left[\log p(\beta|\mathscr{D})\right]$$

Solution:

$$-\log p(\beta|\mathscr{D}) \propto \frac{1}{2\sigma^2} \underbrace{\left(\sum_{i=1}^{n}(y_i - \beta^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)}_{\text{Loss}_{\text{Lasso}}(\beta,\lambda)}$$

$$\underbrace{-n \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - p \log\left(\frac{\lambda}{4\sigma^2}\right)}_{\text{indep. of}\beta}$$

$$\Rightarrow \text{arg}_{\beta} \min[-\log p(\beta|\mathscr{D})] = \text{arg}_{\beta} \min Loss_{Ridge}(\beta, \lambda)$$

- Hence, $\hat{\beta}^{MAP} = \hat{\beta}^{Lasso}$ under prior $\beta_i \sim Lap(0, \frac{2\sigma^2}{\lambda})$ and Gaussian likelihood

Comparison

- In big data problems p can be very large. Least squares will overfit
- Ridge and lasso constrains model complexity by shrinking the parameters
- Lasso sets most of the coefficients to zero
- $\hat{\beta}^{LS}$ is an uniased estimator of β^{true}
- $\hat{\beta}^{Ridge}$ and $\hat{\beta}^{Lasso}$ are biased estimators of β^{true}

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?

$$L(\beta) = p(\mathscr{D}|\beta) \underbrace{=}_{\text{Since } \mathbf{X} \text{ is fixed}} p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \beta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2})$$

$$\sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$$

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

We have

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

Important: Compute the posterior for ridge regression.

Example 5.1

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

Posterior predictive density at test point x:

$$p(y|\mathbf{x}, \mathscr{D}) = \int_{\beta} \underbrace{p(y|\mathbf{x}, \beta)}_{\sim \mathcal{N}(\mathbf{x}^T\beta, \sigma^2)} \; p(\beta, \mathscr{D}) d\beta$$

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

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

Example 3.1 cont.

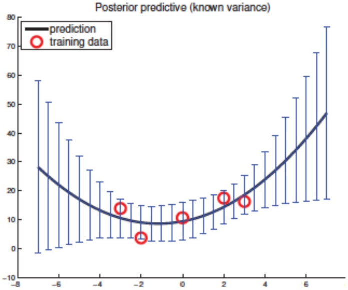


Figure: Murphy, 2012

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

Example 3.1 cont.

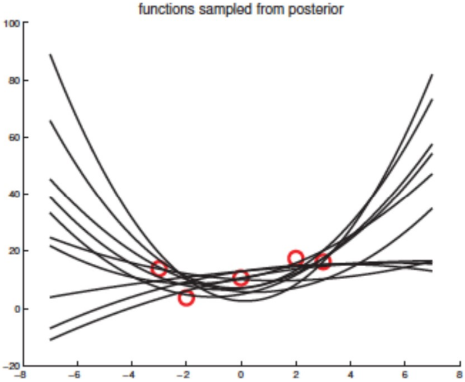


Figure: Murphy, 2012

References

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