Introduction to Statistical Learning

Omid Safarzadeh

January 18, 2022

(ㅁㅏㅓ@ㅏㅓㅌㅏㅓㅌㅏ ㅌ 쒸٩♡

Table of contents

- Likelihood and posterior distribution
 - Computing the posterior
 - Maximum likelihood estimation (MLE)
- Maximum a posteriori (MAP) estimation
 - Posterior mean
 - MAP properties
- Ridge regression from a Bayesian perspective
- 4 Lasso from a Bayesian perspective
 - Comparison
- Bayesian linear regression
- 6 Reference

Omid Safarzadeh Buochisalando Satistical January 18, 2022 2/23

Likelihood and posterior distribution

- $X_i \sim Ber(\theta)$
- Θ: Probability of heads (uncertain value).
- N: Number of coin flips
- \mathcal{D} := { N_1 heads, N_0 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:

$$p_{\mathsf{D}|\Theta}(\mathscr{D}|\theta) = \binom{\mathsf{N}_1 + \mathsf{N}_0}{\mathsf{N}_1} \theta^{\mathsf{N}_1} (1-\theta)^{\mathsf{N}_0}$$

Posterior distribution:

$$\begin{split} p_{\Theta|D}(\theta|\mathscr{D}) &= \frac{p_{D|\Theta}(\mathscr{D}|\theta)p_{\Theta}(\theta)}{p_{D}(\mathscr{D})} \\ &= \frac{\binom{N_1+N_0}{N_1}}{p_{D}(\mathscr{D})} \theta^{N_1} (1-\theta)^{N_0} p_{\Theta}(\theta) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \underbrace{p_{\Theta}(\theta)}_{Q_{\Theta}(\theta)} \end{split}$$

Omid Safarzadeh Busalaguarus Sanskie Harmina January 18, 2022 4

Computing the posterior

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

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

Hence:

$$egin{aligned}
ho_{\Theta|\mathbf{D}}(heta|\mathscr{D}) &\propto heta^{N_1}(1- heta)^{N_0} heta^{s-1}(1- heta)^{b-1} \ &\propto heta^{N_1+s-1}(1- heta)^{N_0+b-1} \ &\Rightarrow \Theta|\mathscr{D} &\sim heta heta(N_1+s,N_0+b) \end{aligned}$$

Omid Safarzadeh Installinston to Structural Resource January 18, 2022 5/23

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:

$$ho(\mathscr{D}| heta)\propto heta^{ extsf{N}_1}(1- heta)^{ extsf{N}_0}$$

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

For the example:

$$\hat{\theta}_{\textit{MLE}} := \textit{arg}_{\theta \in [0,1]} \max \theta^{\textit{N}_1} (1-\theta)^{\textit{N}_0}$$

Omid Safarzadeh Invoduction to Statistical Learning January 18, 2022

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)?



Omid Safarzadeh Indoorgoon op Sterioon Sterioon January 18, 2022

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.

Omid Safarzadeh Introduction to Statistical Learning January 18, 2022

Is MAP a good estimate?

MAP = point estimate (does not measure uncertainty)

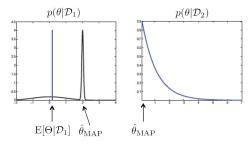


Figure: Murphy, 2012

What does MAP really optimize?

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

10 / 23

Omid Safarzadeh Invocation to Sterilistical Resource January 18, 2022

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

Omid Safarzadeh Introduction to Steakstoch Bermings January 18, 2022

Ridge regression from a Bayesian perspective

Maximum aposteriori estimate of β :

$$\begin{split} \hat{\beta}^{MAP} &= \arg_{\beta} \max p(\beta|\mathscr{D}) \\ &= \arg_{\beta} \max \left[\log p(\beta|\mathscr{D}) \right] \\ &= \arg_{\beta} \min \left[-\log p(\beta|\mathscr{D}) \right] \end{split}$$

Omid Safarzadeh University to Statistical Learning January 18, 2022

Ridge regression from a Bayesian perspective

Solution:

$$-\log p(\beta|\mathscr{D}) \propto \frac{1}{2\sigma^2} \underbrace{\left(\sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2\right)}_{\text{Loss}_{\text{Ridge}}(\beta,\lambda)}$$
$$- \underbrace{n \log(\frac{1}{\sqrt{2\pi\sigma}}) - p \log(\frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma}})}_{\text{indep. of}\beta}$$
$$\Rightarrow \arg_\beta \min[-\log p(\beta|\mathscr{D})] = \arg_\beta \min \text{Loss}_{\text{Ridge}}(\beta,\lambda)$$

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

Omid Safarzadeh Introductivity Statistical Learning January 18, 2022 13

Lasso from a Bayesian perspective

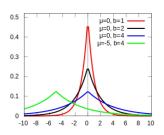
Prior: Laplace prior for β

$$eta_j \sim extit{Lap}(0,rac{2\sigma^2}{\lambda}) \Rightarrow extit{p}(eta_j) = rac{\lambda}{4\sigma^2} ext{exp}(-rac{\lambda}{2\sigma^2}|eta_j|)$$

 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



Omid Safarzadeh Involucion to Statistical Legislatic January 18, 2022

Lasso from a Bayesian perspective

Maximum aposteriori estimate of β :

$$\hat{\beta}^{MAP} = \arg_{\beta} \max \left[\log p(\beta | \mathcal{D}) \right]$$

Solution:

$$-\log p(\beta|\mathscr{D}) \propto \frac{1}{2\sigma^2} \underbrace{(\sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|)}_{\text{Loss}_{\text{Lasso}}(\beta, \lambda)}$$
$$- \underbrace{n \log(\frac{1}{\sqrt{2\pi\sigma}}) - p \log(\frac{\lambda}{4\sigma^2})}_{\text{indep. of}\beta}$$
$$\Rightarrow \arg_{\alpha} \min[-\log p(\beta|\mathscr{D})] = \arg_{\alpha} \min \text{Loss}_{\text{Pidro}}(\beta, \lambda)$$

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

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

Omid Safarzadeh January 18, 2022

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
- ullet \hat{eta}^{LS} is an uniased estimator of eta^{true}
- $\hat{\beta}^{Ridge}$ and $\hat{\beta}^{Lasso}$ are biased estimators of β^{true}



Omid Safarzadeh Introduction to Statistical Learning January 18, 2022

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))
- ullet Bayesian linear regression provides a posterior for eta
- We can specify any prior and likelihood on β !
- But we assume that both prior and likelihood is Gaussian for analytical tractability!

Omid Safarzadeh Installiation to Structural Resolution January 18, 2022 17/23

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 X is observed?

$$L(\beta) = p(\mathcal{D}|\beta) \underbrace{\sum_{\text{Since } \mathbf{X} \text{ is fixed}} p(\mathbf{y}|\mathbf{X}, \beta)}_{\text{Since } \mathbf{X} \text{ is fixed}} = \prod_{i=1}^{n} p(y_{i}|\mathbf{x}_{i}, \beta)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_{i} - \beta^{T}\mathbf{x}_{i})^{2}}{2\sigma^{2}})$$
$$\sim \mathcal{N}(\mathbf{X}\beta, \sigma^{2}\mathbf{I}_{n})$$

• I_n is the n * n identity matrix

Omid Safarzadeh

January 18, 2022

Gaussian posterior

Gaussian prior + Gaussian likelihood ⇒ Gaussian posterior

General formula for Gaussian posterior:

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

We have

$$\bullet \ \Sigma_{\boldsymbol{\beta}|\mathcal{D}}^{-1} = \Sigma_0^{-1} + \boldsymbol{\mathsf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathcal{D}}^{-1} \boldsymbol{\mathsf{X}}$$

$$\bullet \ \mu_{\beta|\mathscr{D}} = \Sigma_{\beta|\mathscr{D}} (\mathbf{X}^{\mathsf{T}} \Sigma_{\mathscr{D}}^{-1} \mathbf{y} + \Sigma_{0}^{-1} \mu_{0})$$

Important: Compute the posterior for ridge regression.

19 / 23

Bayesian linear regression

Example 5.1

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

Posterior predictive density at test point x:

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

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

Omid Safarzadeh Introduction to additional Learning January 18, 2022 20 / 23

Example 3.1 cont.

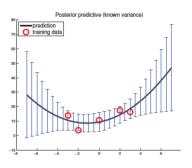


Figure: Murphy, 2012

- 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

Omid Safarzadeh Involucior to Stellerical Learning January 18, 2022 21 / 23

Example 3.1 cont.

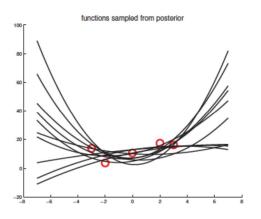


Figure: Murphy, 2012

References

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

Omid Safarzadeh Introduction to Statistical Learning January 18, 2022