Connections between OLS, Lasso, Ridge, MLE, and MAP

1 Introduction

This report is going to show that from statistical perspective, OLS, Lasso, ridge can be explained as a combination of Gaussian/Laplace distribution and MEL/MAP.

Model	Distribution	prior belief
OLS	Gaussian	$\beta \sim U[-\infty, \infty]$
ridge	Gaussian	$\beta \sim N(0, \tau^2)$
lasso	Laplace	$\beta \sim Laplace(0, b)$

2 Understanding Gaussian, Laplace, MLE, and MAP

2.1 Gaussian Distribution

Gaussian distribution has a more familiar name: normal distribution.

1. Notation is: $X \sim N(\mu, \sigma^2)$
2. Its probability distribution function is: $f(x|\mu, \sigma^2) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$

2.2 Laplace Distribution

Laplace distribution is also called double-exponential distribution as there are 2 exponential curves spliced back-to-back. Its pdf does not look as smooth as normal distribution because it has a peak at mean value.

1. Notation is: $X \sim Laplace(\mu, b)$
2. Its probability distribution function is: $f(x|\mu, b) = \frac{e^{-|x-\mu|/b}}{2b}$

2.3 MLE

MLE stands for Maximising Likelihood Estimation. When we want to estimate an unobserved population parameter θ on the basis of observations x and data $D = (x_1, x_2...x_m)$ are independent, sample distribution function is:

$$p(D|\theta) = \prod p(x_i|\theta)$$

likelihood function equals to sample distribution function:

$$L(\theta|D) = p(D|\theta) = \prod p(x_i|\theta)$$

Using ln to make optimisation easier:

$$\theta_{MLE} = argmax_\theta \ \ln L(\theta|D) = argmax_\theta \ \sum \ln p(x_i|\theta)$$

2.4 MAP

MAP stands for Maximise A Posterior, can be seen as MLE plus a regulariser. In MLE case, we treat θ as a given parameter. What if, in a more complex case, the value of θ has a prior distribution $p(\theta)$? We treat θ as a random variable in Bayesian statistic and calculate posterior distribution using Bayes theorem:

$$L(\theta|D) = p(\theta|D) = p(D|\theta)p(\theta) = \prod p(x_i|\theta) \prod p(\theta_i)$$

Using ln optimisation:

$$\theta_{MAP} = argmax_\theta \ln L(\theta|D) = argmax_\theta \sum \ln p(x_i|\theta) + \sum \ln p(\theta_i)$$

MAP equal to MLE when prior distribution is believed to be uniformly distributed in range between negative infinite to positive infinite, $p(\theta_i)$is closing to 0 and thus $\ln p(\theta_i) = 1$ and $\sum \ln p(\theta_i) = C$ a constant value. Therefore, when maximising the likelihood function, C is ignored:

$$\theta_{MAP} = \theta_{MLE} = argmax_\theta \sum \ln p(x_i|\theta)$$

MAP is equal to MLE plus a regulariser, that is: $p(\theta_i) \sim U[-\infty, \infty]$

3 Derivation of OLS , Lasso , ridge

Setting up a simple basic linear model: $Y_i = X_i'\beta + \varepsilon_i$ and considering 3 different strategies:

3.1 Strategy 1 – OLS

1. Assuming that $\varepsilon_i \sim N(0, \sigma^2)$ and $p(\beta_i) \sim U[-\infty, \infty]$, and using MLE/MAP (here makes no statistical difference). Thus $Y_i \sim N(\beta X_i, \sigma^2)$.
2. The likelihood function is:
$$L(\beta) = \prod_{i=1}^{m} f(Y_i) = \frac{\prod_{i=1}^{m} e^{-(Y_i - \beta X_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{\frac{n}{2}}} = \frac{e^{\sum_{i=1}^{m} -(Y_i - \beta X_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{\frac{n}{2}}}$$
3. Maximising ln likelihood:
$$argmax_\beta \ \ln L(\beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (Y_i - \beta X_i)^2 - \frac{m}{2} \ln(2\pi\sigma^2)$$

4. Statistically, maximising the above equation is equal to minimising a OLS regression:
$$argmax_\beta \ \sum_{i=1}^{m} (Y_i - \beta X_i)^2 \ => OLS$$

3.2 Strategy 2 – Ridge

1. Assuming that $\varepsilon_i \sim N(0, \sigma^2)$ and $p(\beta_i) \sim N(0, \tau^2)$ and using MAP.
2. The likelihood function is:
$$L(\beta) = \prod_{i=1}^{m} f(Y_i)p(\beta_i) = \prod_{i=1}^{m} \frac{e^{-(Y_i - \beta X_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{1/2}} \prod_{j=1}^{n} \frac{e^{-\beta_j^2/2\tau^2}}{(2\pi\tau^2)^{1/2}}$$
3. Maximising ln likelihood:
$$argmax_\beta \ \ln L(\beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (Y_i - \beta X_i)^2 - \frac{m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\tau^2} \sum_{j=1}^{n} \beta_j^2 - \frac{n}{2} \ln(2\pi\tau^2)$$
4. Statistically, maximising the above equation is equal to minimising a Ridge regression:
$$argmin_\beta \ \sum_{i=1}^{m} (Y_i - \beta X_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \ => Ridge$$
5. $\lambda[0, \infty]$is a penalty coefficient, if $\lambda = 0$, Ridge = OLS. Ridge is a L2 regulariser, draging all coefficients towards 0.

3.3 Strategy 3 – Lasso

1. Assuming that $\varepsilon_i \sim N(0, \sigma^2)$ and $p(\beta_i) \sim Laplace(0, b)$ and using MAP.

2. The likelihood function is:

$$L(\beta) = \prod_{i=1}^{m} f(Y_i)p(\beta_i) = \prod_{i=1}^{m} \frac{e^{-(Y_i - \beta X_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{1/2}} \prod_{j=1}^{n} \frac{e^{-|\beta_j|/2b}}{2b}$$

3. Maximising ln likelihood:

$$argmax_\beta \ \ln L(\beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i - \beta X_i)^2 - \frac{m}{2}\ln(2\pi\sigma^2) - \frac{1}{2b}\sum_{j=1}^{n}|\beta_j| - n\ln(2b)$$

4. Statistically, maximising the above equation is equal to minimising a Lasso regression:

$$argmin_\beta \ \sum_{i=1}^{m}(Y_i - \beta X_i)^2 + \lambda\sum_{j=1}^{n}|\beta_j| \ => Lasso$$

5. $\lambda[0, \infty]$is penalty coefficient, if $\lambda = 0$, Lasso = OLS . Lasso is a L1 regulariser, it can shrike certain coefficients to 0.

References

[1] Using MAP and Gaussian Distribution to deduce OLS, Lasso and Ridge regressions.(2019) [cnblogs] CJZhao.Simon . https://www.cnblogs.com/jackchen-Net/p/8057726.html

[2] Links and differences between MAP, MLE,and Bayesian. (2018) [csdnblogs].bitcarmanlee . https://blog.csdn.net/bitcarmanlee/article/details/81417151