MLT Week-5

Sherry Thomas
21f3001449

Contents

Abstract

In this week's discussion on Linear Regression, we explore various techniques to minimize Mean Squared Error (MSE) and delve into the concepts of Ridge and Lasso regression. These methods aim to optimize the performance of the linear regression model and improve its predictive power.

Goodness of Maximum Likelihood Estimator for Linear Regression

Given a dataset $\{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$, let $\{y_1, \ldots, y_n\}$ be the labels, where $y_i \in \mathbb{R}$.

$$y|X = w^T x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $w \in \mathbb{R}^d$. Let $\hat{w}_{ML}$ signify the maximum likelihood parameter for linear regression.

$$||w||^2 \leq \theta = w^* = (XX^T)^+ Xy$$

$\therefore$ To measure how good our parameter is, we use the follow:

$$\mathbb{E}[||\hat{w}_{ML} - w||_2^2]$$

This is known as the Mean Squared Error (MSE) and turns out to be equal to

$$\mathbb{E}[||\hat{w}_{ML} - w||_2^2] = \sigma^2 * trace((XX^T)^{-1})$$

Cross-Validation for Minimizing MSE

Let the eigenvalues of XX^T be $\{\lambda_1, \ldots, \lambda_d\}$. Hence the eigenvalues of $(XX^T)^{-1}$ are $\{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d}\}$.

$\therefore$ The MSE is,

$$\mathbb{E}[||\hat{w}_{ML} - w||_2^2] = \sigma^2 \sum_{i=1}^{d} \frac{1}{\lambda_i}$$

Consider the following estimator,

$$\hat{w}_{new} = (XX^T + \lambda I)^{-1} Xy$$

where $\lambda \in \mathbb{R}$ and $I \in \mathbb{R}^{d \times d}$ is the Identity Matrix. Using this we get,

$$trace((XX^T + \lambda I)^{-1}) = \sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}$$

According to the Existence Theorem, $\exists \lambda$ s.t. $\hat{w}_{new}$ has lesser means square error than $\hat{w}_{ML}$.

In practice, we find λ using **cross validation**.

Three main techniques of Cross Validation are:

1. **Training-Validation Split**: The training set is randomly split into training and validation set, usually in the ratio $80 : 20$. From among various λs, we choose the one with gives the least error.
2. **K-Fold Cross Validation**: It is done by dividing the training set into K equally-sized parts, training the model K times on different (K-1) parts, and evaluating it on the remaining part. From among various λs, we choose the one with gives the least average error.
3. **Leave One Out Cross Validation**: It is done by training the model on all but one of the samples in the training set and evaluating it on the left-out sample, repeating this process for each sample in the dataset. From among various λs, we choose the one with gives the least average error.

Bayesian Modeling

An alternate way to understand $\hat{w}_{ML}$ is through Bayesian Modeling.

Let $P(y|X) \sim \mathcal{N}(w^T x, I)$. We use I, the identity matrix, instead of σ^2 for simplicity.

A good choice of prior for w is $\mathcal{N}(0, \gamma^2 I)$, where $\gamma \in \mathbb{R}^d$.

Therefore, we get,

$$\begin{aligned}
P(w|\{(x_1, y_1), \dots, (x_n, y_n)\}) &\propto P(\{(x_1, y_1), \dots, (x_n, y_n)\}|w) * P(w) \\
&\propto \left(\prod_{i=1}^{n} e^{\frac{-(y_i - w^T x_i)^2}{2}}\right) * \left(\prod_{i=1}^{d} e^{\frac{-(w_i - 0)^2}{2\gamma^2}}\right) \\
&\propto \left(\prod_{i=1}^{n} e^{\frac{-(y_i - w^T x_i)^2}{2}}\right) * \left(e^{-\sum_{i=1}^{d} \frac{w_i^2}{2\gamma^2}}\right) \\
&\propto \left(\prod_{i=1}^{n} e^{\frac{-(y_i - w^T x_i)^2}{2}}\right) * e^{\frac{-||w||^2}{2\gamma^2}} \\
\log(P(w|\{(x_1, y_1), \dots, (x_n, y_n)\})) &\propto \frac{-(y_i - w^T x_i)^2}{2} - \frac{||w||^2}{2\gamma^2}
\end{aligned}$$

Differentiating w.r.t. to w and calling it $\widehat{w}_{MAP}, weget$

$$\begin{aligned}\frac{\partial}{\partial w}\log(P(w|\{(x_1,y_1),\dots,(x_n,y_n)\})) &\propto \frac{\partial}{\partial w}\frac{-(y_i-w^Tx_i)^2}{2}-\frac{||w||^2}{2\gamma^2}\\ 0 &= (XX^T)\widehat{w}_{MAP}-Xy+\frac{\widehat{w}_{MAP}}{\gamma^2}\\ \therefore\widehat{w}_{MAP} &= (XX^T+\frac{1}{\gamma^2}I)^{-1}Xy\end{aligned}$$

where $\widehat{w}_{MAP}$ is the Maximum a posteriori Estimate. In practice, the value for $\frac{1}{\gamma^2}$ is acquired using cross validation.

Hence, Maximum a posteriori Estimation for linear regression with a Gaussian Prior $\mathcal{N}(0,\gamma^2 I)$ for w is equivalent to the "new" estimator we used previously.