Assignment1(Part II)

BY XIAONAN PENG

Qishi Intermediate Machine Learning

Ex 3.3, 3.5(Only prove ridge), 3.11 and 3.29 in ESL

Problem. Let $\mathcal{D}=\{X,y\}$ be the collected data, where $X\in\mathbb{R}^{n\times p}$ is the design matrix with full rank and $y\in\mathbb{R}^n$ is the vector of response. Consider the following problem

$$\hat{\beta}=\arg\min\left\{\|b\|_2 \colon b\,\text{minimizes}\,\frac{1}{2n}\|y-Xb\|_2^2\right\}$$

a) Show that the optimal solution of problem (1) is

$$\hat{\beta}=(X^TX)^{-1}X^Ty\,\text{when}\,n\geqslant p$$

$$\hat{\beta}=X^T(XX^T)^{-1}y\,\text{when}\,n<p$$

b) Intialize $\beta^{(0)}=0$, and gradient descent on the least square loss yields

$$\beta^k=\beta^{k-1}+\varepsilon\frac{X^T}{n}(y-X\beta^{(k-1)}),$$

where we take $0<\varepsilon\leqslant\lambda_{\max}(X^TX/n)$(largest eigenvalue). Will the gradient descent converge to the optimal solution given in a)?

c) After rearranging gradient descent, we have

$$\frac{\beta^{(k)}-\beta^{(k-1)}}{\varepsilon}=\frac{X^T}{n}(y-X\beta^{(k-1)}),$$

Setting $\beta(t)=\beta^{(k)}$ at time $t=k\varepsilon$, we have the LHS as the discrete derivative of $\beta(t)$ at time t, which approaches its continuous-time derivative as $\varepsilon\to 0$:

$$\frac{d\beta(t)}{dt}=\frac{X^T}{n}(y-X\beta(t)),$$

over time $t\geqslant 0$, subject to an initial condition $\beta(0)=0$. This is called the **gradient flow differential equation** for the least square problem $\min\frac{1}{2n}\|y-X\beta\|^2$. What is the exact solution path $\beta(t)$ to the above equation for all $t\geqslant 0$.

d) Now consider the ridge regression problem

$$\min_{\beta}\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2$$

What's the degrees of freedom(model flexibility)? You can see that there is not only a shrinkage in weights but also in df as λ increasing. In general, more flexible model has more variance, ridge reduce the model flexibility then reduce the variance.

e) We compare the ridge regression and gradient flow in c).

The ridge regression has form and closed form solution:

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2 \text{ and } \hat{\beta}(\lambda) = (X^TX + n\lambda I)^{-1}X^Ty \tag{1}$$

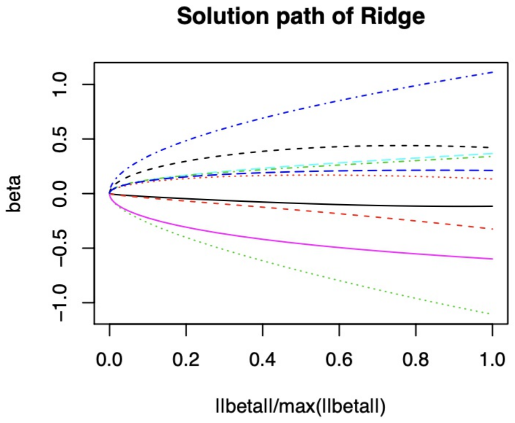
The gradient flow for ordinary least square:

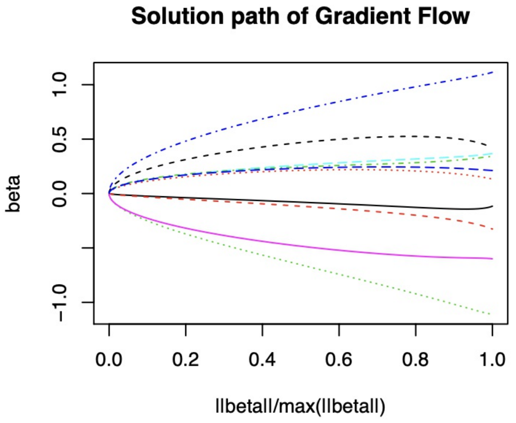
$$\frac{d\beta(t)}{dt} = \frac{X^T}{n}(y - X\beta(t)), \beta(0) = 0 \tag{2}$$

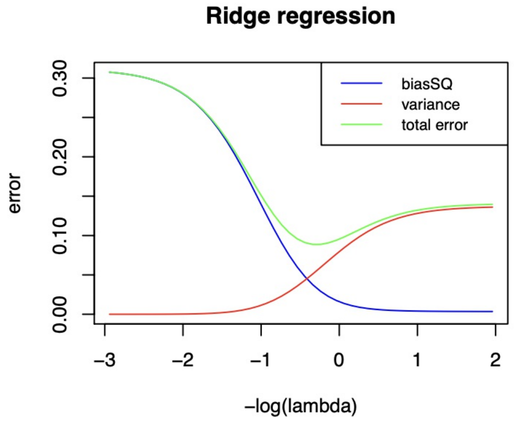
Use simulation to investigate the differences between the solution of (1) and (2). You can use any languages and packages.

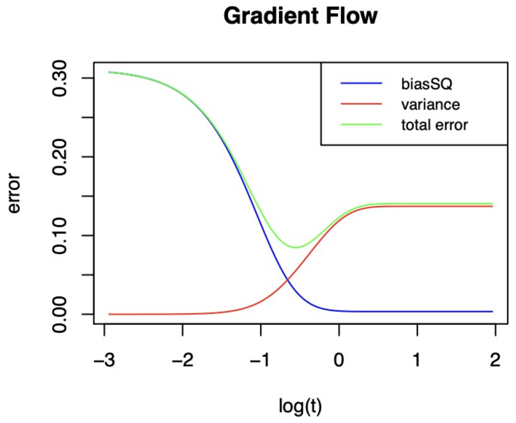
e.g. Comparing these two solution paths. 1. Ridge weights v.s. $\frac{1}{\lambda}$ 2. $\beta(t)$, you will find they are very similar.

The following graphs as reference for you.









Finally, we have a very impressive result: ridge regression with different tunning parameter λ is equivalent to early stopping! $\lambda = 0$, no early stop, $\hat{\beta}(0)$ is the solution of OLS. $\lambda = c$, $\hat{\beta}(c)$ is the solution of OLS with early stop, e.g $\hat{\beta}(c) \approx \beta(\frac{1}{c})$. $\lambda \to \infty$, the solution shrinke to zero, the initial point of gradient flow.

Early stppping is implicit regularization!