Chapter 13

stevenjin8

April 2, 2021

Exercises

Exercise 1

$$\begin{aligned}
\frac{\partial}{\partial w_k}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 &= \frac{\partial}{\partial w_k}\sum(\mathbf{x}_i^T\mathbf{w}-y_i)^2 \\
&= \sum 2(\mathbf{x}_i^T\mathbf{w}-y_i)x_{ik} \\
&= \sum 2(\mathbf{x}_{i,-k}^T\mathbf{w}_{-k}+x_{ik}w_k-y_i)x_{ik} \\
&= \sum 2(\mathbf{x}_{i,-k}^T\mathbf{w}_{-k}+x_{ik}w_k-y_i)x_{ik} \\
&= 2\sum(\mathbf{x}_{i,-k}^T\mathbf{w}_{-k}-y_i)x_{ik}-2\sum x_{ik}^2w_k.
\end{aligned}$$

Setting the above equal to 0 yields

$$\begin{aligned}
\sum(\mathbf{x}_{i,-k}^T\mathbf{w}_{-k}-y_i)x_{ik}-\sum x_{ik}^2w_k &= 0 \\
\mathbf{r}_k^T\mathbf{x}_{:k}-\|\mathbf{x}_{:k}\|_2^2w_k &= 0 \\
\hat{w}_k &= \frac{\mathbf{r}_k^T\mathbf{x}_{:k}}{\|\mathbf{x}_{:k}\|_2^2}
\end{aligned}$$

Exercise 5

I found this question a bit confusing. I think a more straightforward to show that elastic net reduces to lasso is by showing that that the elastic net loss can be rewritten as lasso loss with modified data.

$$\begin{aligned}
J(\mathbf{w}) &= \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2+\lambda_2\|\mathbf{w}\|_2^2+\lambda_1\|\mathbf{w}\|_1 \\
&= \sum_i^N(\mathbf{x}_i^T\mathbf{w}-y_i)^2+\sum_k^D\left(\sqrt{\lambda_2}\mathbf{e}_k^T\mathbf{w}-0\right)^2+\lambda_1\|\mathbf{w}\|_1.
\end{aligned}$$

"Stacking" the sums gives

$$J(\mathbf{w}) = \left\| \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{bmatrix} \mathbf{w} - \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 + \lambda_1 \|\mathbf{w}\|_1.$$

Exercise 6

a. For linear regression, $\hat{w}_k = \frac{c_k}{a_k}$. For lasso, $\hat{w}_k$ is a piecewise linear function of c_k. Finally, for ridge regression, $\hat{w}_k = \frac{c_k}{a_k+2\lambda_2}$. Thus, the dotted line must be lasso. For both ridge and linear regression, $\hat{w}_k$ is a linear function of c_k. But since $\lambda_2 > 0$, the slope for ridge is less steep. Thus, the solid line is linear regression and the dashed line is ridge regression.
b. From figure 13.5, $\lambda_1 = 1$.
c. The slope for the ridge line is $\frac{1}{4}$, while the slope for the linear regression line is $\frac{1}{2}$. Using results from part a, $a_k = 2$ and $a_k + 2\lambda_2 = 4$. Thus, $\lambda_2 = 1$.

Exercise 7

$$p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = \prod_{i=1}^{D} \int_0^1 p(\gamma_i|\pi_i)p(\pi_i|\boldsymbol{\alpha})d\pi_i$$

We can think of the integral as the posterior predictive distribution with no data. Using the results from 3.3.3 and 3.3.4, we find that

$$\begin{aligned} p(\gamma_i = 1|\boldsymbol{\alpha}) &= \int_0^1 p(\gamma_i = 1|\pi_i)p(\pi_i|\boldsymbol{\alpha})d\pi_i \\ &= \frac{\alpha_1}{\alpha_1 + \alpha_2} \end{aligned}$$

Thus,

$$p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = \pi_0^{\|\boldsymbol{\gamma}\|_0}(1-\pi_0)^{D-\|\boldsymbol{\gamma}\|_0}$$

where $\pi_0 = \frac{\alpha_1}{\alpha_1+\alpha_2}$. So, using a Beta prior is the same as using a fixed π_0.

Exercise 8

Using the first hint,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\tau_j^2}\middle| w_j\right] &= \int \frac{1}{\tau_j^2}\frac{\mathcal{N}(w_j|0,\tau_j^2)p(\tau_j^2)}{p(w_j)}d\tau_j^2 \\
&= \frac{1}{p(w_j)}\int \frac{1}{|w_j|}\frac{|w_j|}{2\tau_j^2}\mathcal{N}(w_j|0,\tau_j^2)p(\tau_j^2)d\tau_j^2 \\
&= \frac{1}{p(w_j)}\frac{1}{|w_j|}\int \frac{d}{d|w_j|}\left[\mathcal{N}(w_j|0,\tau_j^2)\right]p(\tau_j^2)d\tau_j^2 \\
&= \frac{1}{p(w_j)}\frac{1}{|w_j|}\frac{d}{d|w_j|}\int \mathcal{N}(w_j|0,\tau_j^2)p(\tau_j^2)d\tau_j^2 \\
&= \frac{1}{|w_j|}\frac{1}{p(w_j)}\frac{d}{d|w_j|}p(w_j) \\
&= \frac{1}{|w_j|}\frac{d}{d|w_j|}\log p(w_j) \\
&= \frac{\pi'(w_j)}{|w_j|}.
\end{aligned}$$

We can further reduce this equation since $p(w_j) = \mathrm{Lap}\left(w_j \middle| 0, \frac{1}{\gamma}\right)$. Also note that $p(w_j)$ is an even function which is why we can mess around with the absolute values.

I found this question interesting for a couple reasons. The arithmetic gymnastics was pretty clever. Another point of interest is the fact that we could have used any prior $p(\tau_j^2)$, not just $p(\tau_j^2) = \mathrm{Ga}(\tau_j^2|1, \frac{\gamma^2}{2})$.

Exercise 9

Recall that for probit regression,

$$p(y|\mathbf{x}) = \Phi\left(\mathbf{w}^T\mathbf{x}\right)^y + \Phi\left(1 - \mathbf{w}^T\mathbf{x}\right)^{1-y}.$$

Thus,

$$\ell(\boldsymbol{\theta}) = \sum\left[y_i \log\left(\mathbf{w}^T\mathbf{x}_i\right) + (1-y_i)\log\left(1-\mathbf{w}^T\mathbf{x}\right)_i\right] - \frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w} + \text{const}.$$

Since τ^2 is independent of $\mathcal{D}$, we can use equation 13.91 to find that

$$\mathbb{E}\left[\frac{1}{\tau^2}\right] = \frac{\gamma}{|w_j|}.$$

Using equation 9.95, the gradient is given by

$$\mathbf{g} = \sum \mathbf{x}_i \frac{\tilde{y}_i\phi(\mathbf{w}^T\mathbf{x}_i)}{\Phi(\tilde{y}_i\mathbf{w}^T\mathbf{x}_i)} - \gamma\,\mathrm{diag}(\mathrm{sign}(w_1),\ldots,\mathrm{sign}(w_D)).$$

We can optimize with any gradient based method. We can see that the regularization term in the gradient "pulls" $\mathbf{w}$ towards $\mathbf{0}$ with constant force γ.