Homework 12

Math 198: Math for Machine Learning

Due Date:
Name:
Student ID:

Instructions for Submission

Please include your name and student ID at the top of your homework submission. You may submit handwritten solutions or typed ones (LaTeX preferred). If you at any point write code to help you solve a problem, please include your code at the end of the homework assignment, and mark which code goes with which problem. Homework is due by start of lecture on the due date; it may be submitted in-person at lecture or by emailing a PDF to both facilitators.

1 Ridge Regression

Consider the linear regression problem in which we seek to fit weights $\mathbf{w}$ given data $\mathbf{X}$, $\mathbf{y}$ and a noise term $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Suppose we have a prior estimate of our parameters' likelihoods; in particular, we assume $w_i \sim \mathcal{N}(0, c)$. Using everything you have learned in this course, derive the optimal values for $\mathbf{w}$ in terms of $\mathbf{X}, \mathbf{y}, \sigma^2$, and c.

We seek to maximize the log-likelihood

$$\log \mathcal{L}(\mathbf{w}, \sigma^2) = \log p(\mathbf{w}) + \sum_{i=1}^{n} \log p(y_i | \mathbf{x_i}; \mathbf{w}, \sigma^2)$$

We first calculate $\log p(\mathbf{w})$:

$$\begin{aligned}
\log p(\mathbf{w}) &= \log \prod_{i=1}^{d} p(w_i) \\
&= \log \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi c}} \exp(-\frac{w_i^2}{2c}) \\
&= \log[(\frac{1}{\sqrt{2\pi c}})^d \exp(-\frac{1}{2c} \sum_{i=1}^{d} w_i^2)] \\
&= -d \log \sqrt{2\pi c} - \frac{1}{2c} \sum_{i=1}^{d} w_i^2
\end{aligned}$$

Recall from note 12 that

$$\sum_{i=1}^{n} \log p(y_i | \mathbf{x_i}; \mathbf{w}, \sigma^2) = -(\frac{n}{2} \log 2\pi + n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x_i}^\top \mathbf{w})^2)$$

So we seek to minimize the negative log likelihood

$$-\log \mathcal{L}(\mathbf{w}, \sigma^2) = d \log \sqrt{2\pi c} + \frac{1}{2c} \sum_{i=1}^{d} w_i^2 + \frac{n}{2} \log 2\pi + n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x_i} \top \mathbf{w})^2$$

Removing the terms which are only constants, we see that this is equivalent to minimizing

$$\frac{1}{2c}\sum_{i=1}^{d} w_i^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^\top \mathbf{w})^2 = \frac{1}{2c}||\mathbf{w}||_2^2 + \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2$$

For simplicity, we will multiply this through by $2\sigma^2$ (this will not change the optimal $\mathbf{w}$) and let $\lambda = \frac{\sigma^2}{c}$, giving us

$$\hat{\mathbf{w}} = \min_w[(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}]$$

Using matrix calculus, we can then derive

$$\begin{aligned} 0 &= -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + 2\lambda\hat{\mathbf{w}} \\ 0 &= -\mathbf{X}^\top\mathbf{y} + \mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}} \\ \mathbf{X}^\top\mathbf{y} &= \mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}} \\ \hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \end{aligned}$$

which matches the formulation we present in note 6, with $\lambda = \frac{\sigma^2}{c}$.