Introduction to Machine Learning

Regularization
Perspectives on Ridge Regression
(Deep-Dive)

Learning goals

- Interpretation of *L2* regularization as row-augmentation
- Interpretation of *L2* regularization as minimizing risk under feature noise

L2 AND ROW-AUGMENTATION

We can also recover the ridge estimator by performing least-squares on a **row-augmented** data set: Let $\tilde{\mathbf{X}} := \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda} \mathbf{I}_p \end{pmatrix}$ and $\tilde{\mathbf{y}} := \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_p \end{pmatrix}$.

With the augmented data, the unreg. least-squares solution $\tilde{\theta}$ is:

$$\begin{aligned} \tilde{\theta} &= \arg\min_{\theta} \sum_{i=1}^{n+p} \left(\tilde{y^{(i)}} - \theta^T \tilde{\mathbf{x}^{(i)}} \right)^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \sum_{j=1}^{p} \left(0 - \sqrt{\lambda} \theta_j \right)^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \lambda \|\theta\|_2^2 \end{aligned}$$

$\Longrightarrow \hat{\theta}_{\text{ridge}}$ is the least-squares solution $\tilde{\theta}$ but using $\tilde{\mathbf{X}}, \tilde{\mathbf{y}}$ instead of $\mathbf{X}, \mathbf{y}$!

This is a sometimes useful "recasting" or "rewriting" for ridge.

L2 AND NOISY FEATURES

Now consider perturbed features $\tilde{x}^{(i)} := \mathbf{x}^{(i)} + \delta^{(i)}$ where $\delta^{(i)} \overset{iid}{\sim} (\mathbf{0}, \lambda \boldsymbol{I}_p)$.
We assume no specifc distribution. Now minimize risk with L2 loss, we define it slightly different than usual, as here our data $\mathbf{x}^{(i)}, y^{(i)}$ are fixed, but we integrate over the random permutations δ:

$$\mathcal{R}(\boldsymbol{\theta}) := \mathbb{E}_\delta\left[\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \tilde{\mathbf{x}}^{(i)})^2\right] = \mathbb{E}_\delta\left[\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top (\mathbf{x}^{(i)} + \delta^{(i)}))^2\right] \quad \Big| \text{expand}$$

$$\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_\delta\left[\sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \delta^{(i)}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \delta^{(i)}\delta^{(i)\top}\boldsymbol{\theta})\right]$$

By linearity of expectation, $\mathbb{E}_\delta[\delta^{(i)}] = \mathbf{0}_p$ and $\mathbb{E}_\delta[\delta^{(i)}\delta^{(i)\top}] = \lambda \boldsymbol{I}_p$, this is

$$\begin{aligned}\mathcal{R}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \mathbb{E}_\delta[\delta^{(i)}](y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \mathbb{E}_\delta[\delta^{(i)}\delta^{(i)\top}]\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda\|\boldsymbol{\theta}\|_2^2\end{aligned}$$

$\Longrightarrow$ Ridge regression on unperturbed features $\mathbf{x}^{(i)}$ turns out to be the same as minimizing squared loss averaged over feature noise distribution!