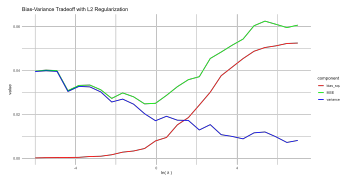


Introduction to Machine Learning

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

PERSPECTIVES ON L_2 REGULARIZATION

We already saw two interpretations of L_2 regularization.

- We know that it is equivalent to a constrained optimization problem:

$$\hat{\theta}_{\text{ridge}} = \arg \min_{\theta} \sum_{i=1}^n \left(y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \lambda \|\theta\|_2^2 = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

For some t depending on λ this is equivalent to:

$$\hat{\theta}_{\text{ridge}} = \arg \min_{\theta} \sum_{i=1}^n \left(y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 \text{ s.t. } \|\theta\|_2^2 \leq t$$

- Bayesian interpretation of ridge regression: For additive Gaussian errors $\mathcal{N}(0, \sigma^2)$ and i.i.d. normal priors $\theta_j \sim \mathcal{N}(0, \tau^2)$, the resulting MAP estimate is $\hat{\theta}_{\text{ridge}}$ with $\lambda = \frac{\sigma^2}{\tau^2}$:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \log[p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)] = \arg \min_{\theta} \sum_{i=1}^n \left(y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \frac{\sigma^2}{\tau^2} \|\theta\|_2^2$$



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 $\tilde{\boldsymbol{\theta}}$ is:

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

$\implies \hat{\boldsymbol{\theta}}_{\text{ridge}}$ is the least-squares solution $\tilde{\boldsymbol{\theta}}$ 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 $\tilde{\mathbf{x}}^{(i)} := \mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}$ where $\boldsymbol{\delta}^{(i)} \stackrel{iid}{\sim} (\mathbf{0}, \lambda \mathbf{I}_p)$.

We assume no specific 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 perturbations $\boldsymbol{\delta}$:



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

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

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

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

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