Machine Learning

Regularization and Feature Selection

Fabio Vandin

November 15^{th}, 2021

Tikhonov regularization

Regularization function: $R(\mathbf{w}) = \lambda||\mathbf{w}||^2$

- $\lambda \in \mathbb{R}, \lambda > 0$
- ℓ_2 norm: $||\mathbf{w}||^2 = \sum_{i=1}^{d} w_i^2$

Therefore the *learning rule* is: pick

$$A(S) = \arg\min_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2 \right)$$

Intuition:

- $||\mathbf{w}||^2$ measures the "complexity" of hypothesis defined by $\mathbf{w}$
- λ regulates the tradeoff between the empirical risk ($L_S(\mathbf{w})$) or overfitting and the complexity ($||\mathbf{w}||^2$) of the model we pick

Ridge Regression

Linear regression with squared loss + Tikhonov regularization
$\Rightarrow$ *ridge regression*

Linear regression with squared loss:

- **given**: training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- **want**: $\mathbf{w}$ which minimizes empirical risk:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

equivalently, find $\mathbf{w}$ which minimizes the *residual sum of squares* $RSS(\mathbf{w})$

$$\mathbf{w} = \arg\min_{\mathbf{w}} RSS(\mathbf{w}) = \arg\min_{\mathbf{w}} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

Linear regression: pick

$$\mathbf{w} = \arg\min_{\mathbf{w}} RSS(\mathbf{w}) = \arg\min_{\mathbf{w}} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

Ridge regression: pick

$$\mathbf{w} = \arg\min_{\mathbf{w}} \left(\lambda ||\mathbf{w}||^2 + \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 \right)$$

RSS: Matrix Form

Let

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}_1 & \cdots \\ \cdots & \mathbf{x}_2 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \mathbf{x}_m & \cdots \end{bmatrix}$$

$\mathbf{X}$: *design matrix*

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$\Rightarrow$ we have that RSS is

$$\sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

Ridge Regression: Matrix Form

Linear regression: pick

$$\arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

Ridge regression: pick

$$\arg\min_{\mathbf{w}} \left(\lambda ||\mathbf{w}||^2 + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})\right)$$

Want to find $\mathbf{w}$ which minimizes
$f(\mathbf{w}) = \lambda||\mathbf{w}||^2 + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$.

How?

Compute gradient $\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}}$ of objective function w.r.t $\mathbf{w}$ and compare it to 0.

$$\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} = 2\lambda\mathbf{w} - 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

Then we need to find $\mathbf{w}$ such that

$$2\lambda\mathbf{w} - 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$$

$$2\lambda \mathbf{w} - 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$$

is equivalent to

$$\left(\lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}\right)\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

Note:

- $\mathbf{X}^T\mathbf{X}$ is positive semidefinite
- $\lambda \mathbf{I}$ is positive definite

$\Rightarrow \lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}$ is positive definite
$\Rightarrow \lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}$ is invertible

Ridge regression solution:
$\mathbf{w} = \left(\lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

Exercise 5

Consider the ridge regression problem $\arg\min_{\mathbf{w}} \lambda||\mathbf{w}||^2 + \sum_{i=1}^{m}(\langle \mathbf{w}, \mathbf{x_i} \rangle - y_i)^2$. Let: h_S be the hypothesis obtained by ridge regression on with training set S; h^* be the hypothesis of minimum generalization error among all linear models.

(A) Draw, in the plot below, a *typical* behaviour of (i) *the training error* and (ii) *the test/generalization error* of h_S as a function of λ.

(B) Draw, in the plot below, a *typical* behaviour of (i) $L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*)$ and (ii) $L_{\mathcal{D}}(h_S) - L_S(h_S)$ as a function of λ.