Machine Learning

Regularization and Feature Selection

Fabio Vandin

November 17^{th}, 2023

Regularized Loss Minimization

Assume h is defined by a vector $\mathbf{w} = (w_1, \ldots, w_d)^T \in \mathbb{R}^d$ (e.g., linear models)

Regularization function $R : \mathbb{R}^d \to \mathbb{R}$

Regularized Loss Minimization (RLM): pick h obtained as

$$\arg\min_{\mathbf{w}} \left(L_S(\mathbf{w}) + R(\mathbf{w}) \right)$$

Intuition: $R(\mathbf{w})$ is a "measure of complexity" of hypothesis h defined by $\mathbf{w}$
$\Rightarrow$ regularization balances between low empirical risk and "less complex" hypotheses

We will see some of the most common regularization function

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

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}_1 & \cdots \\ \cdots & \mathbf{x}_2 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \mathbf{x}_m & \cdots \end{bmatrix} \left.\vphantom{\begin{bmatrix} \cdots \\ \cdots \\ \vdots \\ \cdots \end{bmatrix}}\right\} \text{samples in the training set}$$

$\mathbf{X}$: *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
$f(\mathbf{w}) = \lambda \underbrace{||\mathbf{w}||^2}_{\vec{w}^T \cdot \vec{w}} + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$.

$$||\vec{w}||^2 = \sum_{i=1}^{d} w_i^2 = \vec{w}^T \cdot \vec{w}$$

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?