Machine Learning

Regularization and Feature Selection

Fabio Vandin

November 13^{th}, 2023

Learning Model

- A: learning algorithm for a machine learning task
- S: m i.i.d. pairs $z_i = (x_i, y_i), i = 1, \ldots, m$, with $z_i \in Z = \mathcal{X} \times Y$, generated from distribution $\mathcal{D}$
  $\Rightarrow$ training set available to A to produce $A(S)$;
- $\mathcal{H}$: the hypothesis (or model) set for A
- loss function: $\ell(h, (x, y))$, $\ell : \mathcal{H} \times Z \to \mathbb{R}^+$
- $L_S(h)$: empirical risk or training error of hypothesis $h \in \mathcal{H}$

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

- $L_{\mathcal{D}}(h)$: true risk or generalization error of hypothesis $h \in \mathcal{H}$:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z \in \mathcal{D}}[\ell(h, z)]$$

Learning Paradigms

We would like A to produce $A(S)$ such that $L_{\mathcal{D}}(A(S))$ is *small*, or at least close to the smallest generalization error $L_{\mathcal{D}}(h^*)$ achievable by the "best" hypothesis h^* in $\mathcal{H}$:

$$h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$$

We have seen a *learning paradigm*: Empirical Risk Minimization

"$A(S)$ should be the hyp. of smallest empirical risk $L_S(h)$"

We will now see another learning paradigm...

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

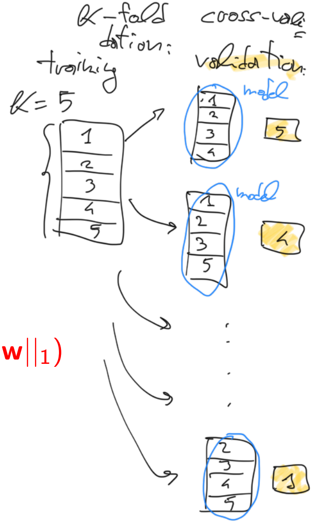
Regularized Loss Minimization (RLM): pick h obtained as

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

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

ℓ_1 Regularization



Regularization function: $R(\mathbf{w}) = \lambda ||\mathbf{w}||_1$

- $\lambda \in \mathbb{R}, \lambda > 0$
- ℓ_1 norm: $||\mathbf{w}||_1 = \sum_{i=1}^{d} |w_i|$

Therefore the *learning rule* is: pick

$$A(S) = \arg\min_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda ||\mathbf{w}||_1 \right)$$

Intuition:

- $||\mathbf{w}||_1$ 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 ($||\mathbf{w}||_1$) of the model we pick

LASSO

Linear regression with squared loss + ℓ_1 regularization $\Rightarrow$ *LASSO* (*least absolute shrinkage and selection operator*)

LASSO: pick

$$\mathbf{w} = \arg\min_{\mathbf{w}} \underbrace{\lambda||\mathbf{w}||_1}_{\ell_1 \text{ regularization}} + \overbrace{\underbrace{\sum_{i=1}^{m}(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2}_{L_S(h_{\vec{w}})}}^{(= m\, L_S(h_{\vec{w}}))}$$

How?

Notes:

- no closed form solution!
- ℓ_1 norm is a convex function and squared loss is convex $\Rightarrow$ problem can be solved efficiently! (true for every convex loss function)

LASSO and Sparse Solutions: Example

(Equivalent) one dimensional regression problem with squared loss:

$\mathcal{X} = \mathbb{R}$

$$\arg\min_{w \in \mathbb{R}} \left(\frac{1}{2m} \sum_{i=1}^{m} (x_i w - y_i)^2 + \lambda |w| \right)$$

Is equivalent to:

$$\arg\min_{w \in \mathbb{R}} \left(\frac{1}{2} \left(\frac{1}{m} \sum_{i=1}^{m} x_i^2 \right) w^2 - \left(\frac{1}{m} \sum_{i=1}^{m} x_i y_i \right) w + \lambda |w| \right)$$

Assumptions $\rightarrow$ Assume for simplicity that $\frac{1}{m} \sum_{i=1}^{m} x_i^2 = 1$, and let $\sum_{i=1}^{m} x_i y_i = \langle \mathbf{x}, \mathbf{y} \rangle$.

Then the optimal solution is

$$w = \text{sign}(\langle \mathbf{x}, \mathbf{y} \rangle) \underbrace{[\langle \mathbf{x}, \mathbf{y} \rangle / m - \lambda]_+}_{}$$

where $[a]_+ =^{(def)} \max\{a, 0\}$.

$\lambda >> 0 \longrightarrow w = 0$

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:

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

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

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

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

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