Machine Learning

Regularization and Feature Selection

Fabio Vandin

November 5, 2020

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 paradigms*: Empirical Risk Minimization

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

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: $R(\mathrm{w}) = \lambda ||\mathrm{w}||_1$

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

Therefore the *learning rule* is: pick

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

Intuition:

- $||\mathrm{w}||_1$ measures the "complexity" of hypothesis defined by w
- λ regulates the tradeoff between the empirical risk ($L_S(\mathrm{w})$) or overfitting and the complexity ($||\mathrm{w}||_1$) of the model we pick

LASSO

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

LASSO: pick

$$\mathrm{w} = \arg\min_{\mathrm{w}} \lambda ||\mathrm{w}||_1 + \sum_{i=1}^{m} (\langle \mathrm{w}, \mathrm{x}_i \rangle - y_i)^2$$

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)

Tikhonov regularization

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

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

Therefore the *learning rule* is: pick

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

Intuition:

- $||\mathrm{w}||^2$ measures the "complexity" of hypothesis defined by w
- λ regulates the tradeoff between the empirical risk ($L_S(\mathrm{w})$) or overfitting and the complexity ($||\mathrm{w}||^2$) 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 $S = ((\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_m, y_m))$, with $\mathrm{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- **want**: w which minimizes empirical risk:

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

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

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

Linear regression: pick

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

Ridge regression: pick

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

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

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

Ridge regression: pick

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

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

How?

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

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

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

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

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

is equivalent to

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

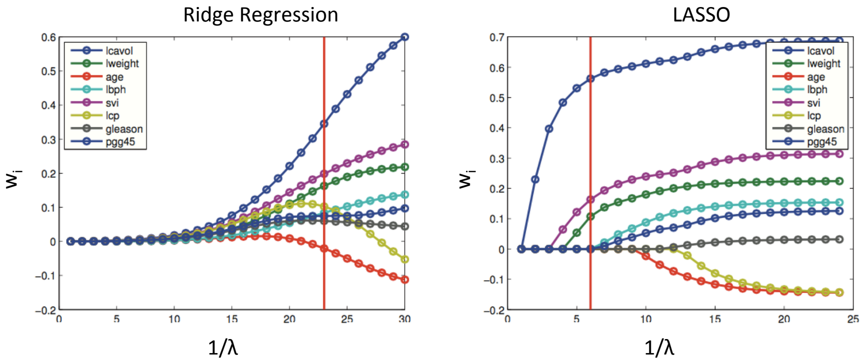
Note:

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

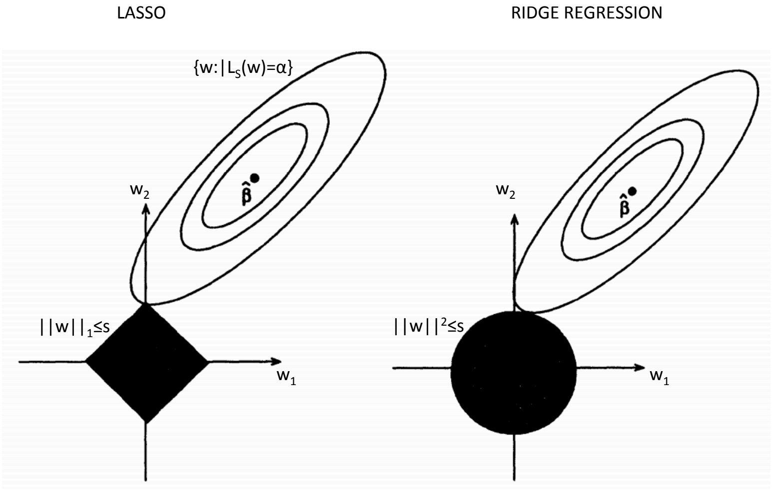
$\Rightarrow \lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}$ is positive definite
$\Rightarrow \lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}$ is invertible

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

Ridge Regression vs LASSO: Sparsity of Solutions



Ridge Regression vs LASSO



ℓ_1 regularization performs a sort of **feature selection**

Exercise 5

Consider the ridge regression problem $\arg\min_{\mathrm{w}} \lambda ||\mathrm{w}||^2 + \sum_{i=1}^{m}(\langle \mathrm{w}, \mathrm{x_i} \rangle - y_i)^2$. 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.