Machine Learning

Regularization and Feature Selection

Fabio Vandin November 13, 2017
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 \( z_i \in Z = \mathcal{X} \times \mathcal{Y} \), 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 \( h \in \mathcal{H} \)

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

• \( L_D(h) \): true risk or generalization error of hypothesis \( h \in \mathcal{H} \):

\[
L_D(h) = \mathbb{E}_{z \in \mathcal{D}}[\ell(h, z)]
\]
Learning Paradigms

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

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

We have seen two learning paradigms:

- Empirical Risk Minimization
- Structural Risk Minimization $\Rightarrow$ Minimum Description Length
• Empirical Risk Minimization (ERM): pick

\[ A(S) \in \arg \min_{h \in \mathcal{H}} L_S(h) \]

⇒ guarantees learning for PAC learnable \( \mathcal{H} \)

• Minimum Description Length: pick

\[ A(S) \in \arg \min_{h \in \mathcal{H}} \left( L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}} \right) \]

where \(|h|\) is the length of the description of \( h \).
⇒ considers trade-off between training error and complexity

We now see another learning paradigm.
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

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

**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: \( R(w) = \lambda \|w\|^2 \)

- \( \lambda \in \mathbb{R}, \lambda > 0 \)
- \( \ell_2 \) norm: \( \|w\|^2 = \sum_{i=1}^{d} w_i^2 \)

Therefore the learning rule is: pick

\[
A(S) = \arg \min_w \left( L_S(w) + \lambda \|w\|^2 \right)
\]

Intuition:
- \( \|w\|^2 \) measures the “complexity” of hypothesis defined by \( w \)
- \( \lambda \) regulates the tradeoff between the empirical risk \( (L_S(w)) \) or overfitting and the complexity \( (\|w\|^2) \) of the model we pick
Ridge Regression

Linear regression with squared loss + Tikhonov regularization
⇒ ridge regression

Linear regression with squared loss:

• given: training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$

• want: $w$ which minimizes empirical risk:

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

equivalently, find $w$ which minimizes the residual sum of squares $RSS(w)$

$$w = \arg \min_w RSS(w) = \arg \min_w \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2$$
Linear regression: pick

\[ w = \arg \min_w \text{RSS}(w) = \arg \min_w \sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2 \]

Ridge regression: pick

\[ w = \arg \min_w \lambda \|w\|^2 + \sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2 \]
RSS: Matrix Form

Let

\[ X = \begin{bmatrix} \cdots & x_1 & \cdots \\ \cdots & x_2 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & x_m & \cdots \end{bmatrix} \]

\[ X: \text{ design matrix} \]

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \]

\( \Rightarrow \) we have that RSS is

\[
\sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2 = (y - Xw)^T (y - Xw)
\]
Linear regression: pick

$$\arg \min_w (y - Xw)^T (y - Xw)$$

Ridge regression: pick

$$\arg \min_w \lambda \|w\|^2 + (y - Xw)^T (y - Xw)$$
Want to find $w$ which minimizes

$$f(w) = \lambda ||w||^2 + (y - Xw)^T (y - Xw).$$

How?

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

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

Then we need to find $w$ such that

$$2\lambda w - 2X^T (y - Xw) = 0$$
\[ 2\lambda w - 2X^T(y - Xw) = 0 \]

is equivalent to

\[ \left( \lambda I + X^TX \right) w = X^Ty \]

Note:

- \( X^TX \) is positive semidefinite
- \( \lambda I \) is positive definite

\[ \Rightarrow \lambda I + X^TX \] is positive definite

\[ \Rightarrow \lambda I + X^TX \] is invertible

Ridge regression solution:

\[ w = \left( \lambda I + X^TX \right)^{-1} X^Ty \]
We’ll show: Tikhonov regularization makes the learner *stable* w.r.t. small perturbations of the training set, which in turn leads to small bounds on generalization error.

Informally: an algorithm \( A \) is *stable* if a small change of the training data (i.e., its input) \( S \) will lead to a small change of its output hypothesis.

Questions:
- what is a “small change of the training data”?
- what is a “small change of its output hypothesis”?
Stability

• “small change of the training data” = replace one sample!

Given $S = (z_1, \ldots, z_m)$ and additional example (i.e., pair instance label/target) $z'$ let $S(i) = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$

• “small change of its output hypothesis” = on-average-replace-one-stable (OAROS)

**Definition**

Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm $A$ is on-average-replace-one-stable OAROS with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$:

$$E_{(S,z') \sim \mathcal{D}^{m+1}, i \sim U(m)}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \epsilon(m)$$
**Proposition**

If algorithm $A$ is OAROS with rate $\epsilon(m)$ then:

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_D(A(S)) - L_S(A(S))] \leq \epsilon(m)$$

**Proof.**

Since $S$ and $z'$ are both drawn i.i.d. from $\mathcal{D}$, we have that for every $i$:

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_{S,z'}[\ell(A(S), z')] = \mathbb{E}_{S,z'}[\ell(A(S^{(i)}), z_i)].$$

On the other hand we have:

$$\mathbb{E}_S[L_S(A(S))] = \mathbb{E}_{S,i}[\ell(A(S), z_i)].$$

The proof follows from the definition of stability.
Tikhonov Regularization is a Stabilizer

**Proposition**

Assume the loss function is convex and $\rho$-Lipschitz continuous. Then, the RLM rule with regularizer $\lambda \|w\|^2$ is OAROS with rate $\frac{2\rho}{\lambda m}$. It follows that for the RLM rule:

$$\mathbb{E}_{S \sim D^m}[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho}{\lambda m}$$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is $\rho$-Lipschitz continuous if for every $w_1, w_2 \in \mathbb{R}^d$ we have that $\|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|$
The Fitting-Stability Tradeoff

Note that

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_D(A(S)) - L_S(A(S))]$$

Notes:

- $\mathbb{E}_S[L_S(A(S))]$: how well $A$ fits the training set
- $\mathbb{E}_S[L_D(A(S)) - L_S(A(S))] = \text{overfitting}$, bounded by stability of $A$
- in Tikhonov regularization, $\lambda$ controls tradeoff between the two terms

Questions:

- how do $L_S(A(S))$ and $||A(S)||^2 = ||w||^2$ vary as a function of $\lambda$?
- how may $\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]$ change as a function of $\lambda$?

How do set $\lambda$?
Using the fitting-stability tradeoff decomposition and the result for convex, Lipschitz losses we can prove that knowing the properties (e.g., $\rho$ in $\rho$-Lipschitz continuity, etc.) one can pick $\lambda$ to guarantee that

$$\mathbb{E}_S[L_D(A(S))] \leq \min_{w \in \mathcal{H}} L_D(w) + \sqrt{\frac{c}{m}}$$

where $c > 0$ depends on the parameters of the loss function.

**Question**: how do we pick $\lambda$ in practice?

**Answer**: validation!