ISML-II: Machine Learning **Spring 2014**

Lecture 7- Regularized Least Squares

Lecturer: Lorenzo Rosasco *Scribe: Lorenzo Rosasco*

In this class we introduce a class of learning algorithms based Tikhonov regularization, a.k.a. penalized empirical risk minimization and regularization. In particular, we focus on the algorithm defined by the square loss.

7.1 Regularized Least Squares

We consider the following algorithm

$$\min_{w\in\mathbb{R}^D} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i))^2 + \lambda w^T w, \quad \lambda \geq 0. \tag{7.1}$$

A motivation for considering the above scheme is to view the empirical error

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i))^2,$$

as a proxy for the expected error

$$\int dxdyp(x,y)(y - w^T x))^2.$$

The term $w^T w$ is a regularizer and help preventing overfitting.

The term $w^T w = \|w\|^2$ is called regularizer and controls the stability of the solution. The parameter λ balances the error term and the regularizer. Algorithm (7.1) is an instance of Tikhonov regularization, also called penalized empirical risk minimization. We have implicitly chosen the space of possible solution, called the hypotheses space, to be the space of linear functions, that is

$$H = \{f : \mathbb{R}^D \to \mathbb{R} \ : \ \exists w \in \ \mathbb{R}^D \text{ such that } f(x) = x^T w, \ \forall x \in \mathbb{R}^D\},$$

so that finding a function f_w reduces to finding a vector w. As we will see in the following, this seemingly simple example will be the basis for much more complicated solutions.

7.2 Computations

In this case it is convenient to introduce the n times D matrix X_n, where the rows are the input points, and the n by 1 vector Y_n where the entries are the corresponding outputs. With this notation

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i))^2 = \frac{1}{n}\|Y_n - X_n w\|^2.$$

A direct computation shows that the gradient with respect to w of the empirical risk and the regularizer are respectively

$$-\frac{2}{n}X_n^T(Y_n - X_n w), \quad \text{and}, \quad 2w.$$

Then, setting the gradient to zero, we have that the solution of regularized least squares solves the linear system

$$(X_n^T X_n + \lambda n I)w = X_n^T Y_n.$$

Several comments are in order. First, several methods can be used to solve the above linear systems, Choleski decomposition being the method of choice, since the matrix $X_n^T X_n + \lambda I$ is symmetric and positive definite. The complexity of the method is essentially $O(nd^2)$ for training and $O(d)$ for testing. The parameter λ controls the *invertibility* of the matrix $(X_n^T X_n + \lambda n I)$.

7.3 Interlude: Linear Systems

Consider the problem

$$Ma = b,$$

where M is a D by D matrix and a, b vectors in $\mathbb{R}^D$. We are interested in determing a satisfying the above equation given M, b. If M is invertible, the solution to the problem is

$$a = M^{-1}b.$$

- If M is a diagonal $M = diag(\sigma_1, \ldots, \sigma_D)$ where $\sigma_i \in (0, \infty)$ for all $i = 1, \ldots, D$, then

$$M^{-1} = diag(1/\sigma_1, \ldots, 1/\sigma_D), \quad (M + \lambda I)^{-1} = diag(1/(\sigma_1 + \lambda), \ldots, 1/(\sigma_D + \lambda)$$

- If M is symmetric and positive definite, then considering the eigendecomposition

$$M^{-1} = V\Sigma V^T, \quad \Sigma = diag(\sigma_1, \ldots, \sigma_D), \; VV^T = I,$$

then

$$M^{-1} = V\Sigma^{-1}V^T, \quad \Sigma^{-1} = diag(1/\sigma_1, \ldots, 1/\sigma_D),$$

and

$$(M + \lambda I)^{-1} = V\Sigma_\lambda V^T, \quad \Sigma_\lambda = diag(1/(\sigma_1 + \lambda), \ldots, 1/(\sigma_D + \lambda)$$

The ratio σ_D/σ_1 is called the condition number of M.

7.4 Dealing with an Offset

When considering linear models, especially in relatively low dimensional spaces, it is interesting to consider an offset, that is $w^T x + b$. We shall ask the question of how to estimate b from data. A simple idea is to simply augment the dimension of the input space, considering

$\tilde{x} = (x, 1)$ and $\tilde{w} = (w, b)$. While this is fine if we do not regularize, if we do then we still tend to prefer linear functions passing through the origin, since the regularizer becomes

$$\|\tilde{w}\|^2 = \|w\|^2 + b^2.$$

In general we might not have reasons to believe that the model should pass through the origin, hence we would like to consider an offset and still regularize considering only $\|w\|^2$, so that the offset is not penalized. Note that the regularized problem becomes

$$\min_{(w,b)\in\mathbb{R}^{D+1}} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b)^2 + \lambda\|w\|^2.$$

The solution of the above problem is particularly simple when considering least squares. Indeed, in this case it can be easily proved that a solution w^*, b^* of the above problem is given by

$$b^* = \bar{y} - \bar{x}^T w^*$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and w^* solves

$$\min_{w\in\mathbb{R}^{D+1}} \frac{1}{n}\sum_{i=1}^{n}(y_i^c - w^T x_i^c)^2 + \lambda\|w\|^2.$$

where $y_i^c = y - \bar{y}$ and $x_i^c = x - \bar{x}$ for all $i = 1, \ldots, n$.