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

Lecture 8- Regularized Least Squares Classification

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.

While least squares are often associated to regression problem, we next discuss theyr interpretation in the context of binary classification and discuss an extension to multi class classification.

8.1 Nearest Centroid Classifier

Let's consider a classification problem and assume that there is an equal number of point for class 1 and -1. Recall that the nearest centroid rule is given by

$$\text{sign}h(x), \quad h(x) = \|x - m_{-1}\|^2 - \|x - m_1\|^2$$

where

$$m_1 = \frac{2}{n} \sum_{i \mid y_i = 1} x_i, \quad m_{-1} = \frac{2}{n} \sum_{i \mid y_i = -1} x_i.$$

It is easy to see that we can write,

$$h(x) = x^T w + b, \quad w = m_1 - m_{-1}, \quad b = -(m_1 - m_{-1})^T m,$$

where

$$m = m_1 + m_{-1} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

In a compact notation we can write,

$$h(x) = (x - m)^T (m_1 - m_{-1}).$$

8.2 RLS for Binary Classification

If we consider an offset, the classification rule given by RLS is

$$\text{sign}f(x), \quad f(x) = x^T w + b,$$

where

$$b = -m^T w,$$

since $\frac{1}{n} \sum_{i=1}^{n} y_i = 0$ by assumption, and

$$w = (\overline{X}_n^T \overline{X}_n + \lambda n I)^{-1} \overline{X}_n^T Y_n = (\frac{1}{n} \overline{X}_n^T \overline{X}_n + \lambda I)^{-1} \frac{1}{n} \overline{X}_n^T Y_n,$$

with $\overline{X}_n$ the *centered* data matrix having rows $x_i - m$, $i = 1, \ldots, m$.

It is easy to show a connection between the RLS classification rule and the nearest centroid rule. Note that,

$$\frac{1}{n}\overline{X}_N^T Y_n = \frac{1}{n} X_N^T Y_n = m_1 - m_{-1},$$

so that, if we let $C_\lambda = \frac{1}{n}\overline{X}_n^T \overline{X}_n + \lambda I$

$$b = -m^T C_\lambda^{-1}(m_1 - m_{-1})$$

and

$$f(x) = (x - m)^T C_\lambda^{-1}(m_1 - m_{-1})$$

If λ is large then $(\frac{1}{n} X_n^T X_n + \lambda I) \sim \lambda I$, and we see that

$$f(x) \sim \frac{1}{\lambda} h(x) \Leftrightarrow \text{sign} f(x) = \text{sign} h(x).$$

If λ is small $C_\lambda \sim C = \frac{1}{n}\overline{X}_n^T \overline{X}_n$, the inner product $x^T w$ is replaced with a new inner product $(x - m)^T C^{-1}(x - m)$. The latter is the so called Mahalanobis distance. If we consider the eigendecomposition of $C = V \Sigma V^T$ we can better understand the effect of the new inner product. We have

$$f(x) = (x - m)^T V \Sigma^{-1} \lambda^{-1} V^T (m_1 - m_{-1}) = (\tilde{x} - \tilde{m})^T (\tilde{m}_1 - \tilde{m}_{-1}),$$

where $\tilde{u} = \Sigma^{1/2} V^T u$. The data are rotated and then stretched in directions where the eigenvalues are small.

8.3 RLS for Multiclass Classification

RLS can be adapted to problem with $T > 2$ classes considering

$$(X_n^T X_n + \lambda n I) W = X_n^T Y_n. \tag{8.1}$$

where W is a D by T matrix, and Y_n is a n by T matrix where the i-th column has entry 1 if the corresponding input belongs to the i-th class and -1 otherwise. If we let w_t, $t = 1, \ldots, T$, denote the columns of W then the corresponding classification rule $c : X \to \{1, \ldots, T\}$ is

$$c(x) = \arg \max_{t=1,\ldots,T} x^T W^t$$

The above scheme can be seen as a reduction scheme from multi class to a collection of binary classification problems. Indeed, the solution of 8.1 can be shown to solve the minimization problem

$$\min_{W^1,\ldots,W^T} \sum_{t=1}^{T} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i^t - x_i^T W^t)^2 + \lambda \|W^t\|^2\right).$$

where $y_i^t = 1$ if the x_i belong to class t and $y_i^t = -1$, otherwise. The above minimization can be done separately for all w_i, $i = 1, \ldots, T$. Each minimization problem can be interpreted as performing a *"one vs all"* binary classification.