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

Lecture 13- Regularization Networks

Lecturer: Lorenzo Rosasco *Scribe: Lorenzo Rosasco*

In this class we introduce a class of learning algorithms based on Tikhonov regularization, a.k.a. penalized empirical risk minimization and regularization. In particular, we study common computational aspects of these algorithms introducing the so called representer theorem.

13.1 Empirical Risk Minimization

Among different approaches to design learning algorithms, empirical risk minimization (ERM) is probably the most popular one. The general idea behind this class of methods is to consider the empirical error

$$\hat{\mathcal{E}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)),$$

as a proxy for the expected error

$$\mathcal{E}(f) = \mathbb{E}[\ell(y, f(x))] = \int dxdyp(x,y)\ell(y, f(x)).$$

Recall that ℓ is a loss function and measure the price we pay predicting $f(x)$ when in fact the right label is y. Also, recall that the expected error cannot be directly computed since the data distribution is fixed but unknown.

In practice, to turn the above idea into an actual algorithm we need to fix a suitable hypotheses space H on which we will minimize $\hat{\mathcal{E}}$.

13.2 Hypotheses Space

The hypotheses space should be such that computations are feasible, at the same time it should be *rich* since the complexity of the problem is not known a priori. As we have seen, the simplest example of hypotheses space is the space of linear functions, that is

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

Each function f is defined by a vector w and we let $f_w(x) = x^T w$. We have also seem how we can vastly extend the class of functions we can consider by introducing a feature map

$$\Phi : \mathbb{R}^D \to \mathbb{R}^p,$$

where typically $p \gg D$, and considering functions of the form $f_w(x) = \Phi(x)^T w$. We have also seen how this model can be pushed further considering so called reproducing kernels

$$K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$$

that is symmetric and positive definite functions, implicitly defining a feature map via the equation

$$\Phi(x)^T\Phi(x') = K(x,x').$$

If the hypotheses space is rich enough, solely minimizing the empirical risk is not enough to ensure a generalizing solution. Indeed, simply solving ERM would lead to estimators which are highly dependent on the data and could overfit. Regularization is a general class of techniques that allow to restore stability and ensure generalization.

13.3 Tikhonov regularization and Representer Theorem

We consider the following *Tikhonov* regularization scheme,

$$\min_{w\in\mathbb{R}^D} \hat{\mathcal{E}}(f_w) + \lambda\|w\|^2. \tag{13.1}$$

The above scheme describes a large class of methods sometimes called Regularization Networks. The term $\|w\|^2$ is called regularizer and controls the stability of the solution. The parameter λ balances the error term and the regularizer.

Different classes of methods are induced by the choice of different loss functions In the following, we will see common aspects and differences in considering different loss functions.

There is no general computation scheme to solve problems of the form (13.1), and the actual solution for each algorithm depend on the considered loss function. However, we show next that for linea function the solution of problem (13.1) can always be written as

$$w = X_n^t c, \quad f(x) = \sum_{i=1}^n x_i^T x c_i \tag{13.2}$$

where X_n is the n by D data matrix and $c = (c_1, \dots, c_n)$. This allows on the one hand to reduce computational complexity when $n \ll D$, or $n \ll p$ in the case of feature map.

13.3.1 Representer Theorem for General Loss Functions

Here we discuss the general proof of the representer theorem for loss function other than the square loss.

- The vectors of the form (13.2) form a linear subspace $\widehat{W}$ of $\mathbb{R}^D$. Hence for every $v \in \mathbb{R}^D$ we have the decomposition $w = \hat{w} + \hat{w}^\perp$, where $\hat{w} \in \widehat{W}$ and $\hat{w}^\perp$ belongs to the space $\widehat{W}^\perp$ of vectors orthogonal to those in $\widehat{W}$, i.e.

$$\hat{w}^T\hat{w}^\perp = 0. \tag{13.3}$$

- The following is the key observation: for all $i = 1, \dots, n$ $x_i \in \widehat{W}$, so that

$$f_w(x_i) = x_i^T w = x_i^T(\hat{w} + \hat{w}^\perp) = x_i^T\hat{w}.$$

It follows that the empirical error depends only on $\hat{w}$!

- For the regularizer we have

$$\|w\|^2 = \|\hat{w} + \hat{w}^\perp\|^2 = \|\hat{w}\|^2 + \|\hat{w}^\perp\|^2,$$

because of (13.3). Clearly the above expression is minimized if we take $\hat{w}^\perp = 0$.

The theorem is hence proved, the first term in (13.1) depends only on vector of the form (13.2) and the same form is the best to minimize the second term

13.4 Loss Functions and Target Functions

It is useful to recall that different loss function might define different goal via the corresponding target functions.

A simple calculation shows what is the target function corresponding to the square loss. Recall that the target function minimize the expected squared loss error

$$\mathcal{E}(f) = \int p(x,y)dxdy(y - f(x))^2 = \int p(x)dx \int p(y|x)dy(y - f(x))^2.$$

To simplify the computation we let

$$f^*(x) = \arg\min_{a \in \mathbb{R}} \int p(y|x)dy(y-a)^2,$$

for all $x \in X$. It is easy to see that the solution is given by

$$f^*(x) = \int dyp(y|x)y.$$

In classification

$$f^*(x) = 2p - 1, \quad p = p(1|x),$$

which justify taking the sing of f.

Similarly we can derive the target function of the logistic loss function,

$$f^*(x) = \arg\min_{a \in \mathbb{R}} \int p(y|x)dy \log(1 + e^{-ya}) = \arg\min_{a \in \mathbb{R}} p \log(1 + e^{-a}) + (1-p)\log(1 + e^{a}).$$

We can simply take the derivative and set it equal to zero,

$$p\frac{-e^{-a}}{(1+e^{-a})} + (1-p)\frac{e^a}{(1+e^a)} = -p\frac{1}{(1+e^{-a})} + (1-p)\frac{e^a}{(1+e^a)},$$

so that

$$p = \frac{e^a}{(1+e^a)} \Longrightarrow a = \log\frac{p}{1-p}$$

A similar computation allows to