MLCC: Machine Learning Crash Course **Spring 2014**

Lecture 3- Regularization Network I: Linear Case

Lecturer: F.Odone–L. Rosasco

In this class we introduce a class of learning algorithms based on empirical risk minimization and regularization.

3.1 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 in an actual algorithm we need to fix a suitable hypotheses space H on which we will minimize $\hat{\mathcal{E}}$.

3.2 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. 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 $f_w(x) = x^T w$. As we will see in the following, this seemingly simple example will be the basis for much more complicated solutions.

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.

3.3 Tikhonov regularization

We consider the following *Tikhonov* regularization scheme,

$$\min_{w \in \mathbb{R}^D} \hat{\mathcal{E}}(f_w) + \lambda \|w\|^2. \tag{3.1}$$

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

3.4 Regularized Least Squares

Consider $\ell(y, f_w(x)) = (y - f_w(x))^2$.The corresponding regularized empirical risk minimization problem defines the regularized least squares algorithm, a.k.a. ridge regression. 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

$$\hat{\mathcal{E}}(f_w) = \frac{1}{n}\|Y_n - X_n w\|^2.$$

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

3.5 Regularized Logistic Regression

Consider $\ell(y, f_w(x)) = \log(1 + e^{-y f_w(x)})$, namely the logistic loss function. The corresponding regularized empirical risk minimization problem is called regularized logistic regression. Its solution can be computed via gradient descent. For $w_0 = 0$, let

$$w_t = w_{t-1} - \frac{\gamma}{n} \left(\sum_{i=1}^{n} \frac{-y_i x_i e^{-y_i x_i^T w_{t-1}}}{1 + e^{-y_i x_i^T w_{t-1}}} + 2\lambda w_{t-1} \right)$$

for $t = 1, \dots T$, where

$$\frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_i e^{-y_i x_i^T w}}{1+e^{-y_i x_i^T w}} + 2\lambda w = \nabla(\hat{\mathcal{E}}(f_w) + \lambda \|w\|^2)$$

The solution of logistic regression can be shown to have probabilistic interpretation, in fact it can be derived from the following model

$$p(1|x) = \frac{e^{x^T w}}{1+e^{x^T w}}$$

where the right hand side is called logistic function. This latter observation can be used to deduce a confidence from the on each prediction of the logistic regression estimator.

3.6 Support Vector Machines

Consider the so called *hinge* loss $\ell(y, f_w(x)) = |1 - y f_w(x)|_+$, where $|a|_+ = a$, if $a > 0$ and $|a|_+ = 0$, otherwise. The corresponding regularized empirical risk minimization problem defines the Support Vector Machines algorithm. The following formulation is sometimes considered

$$\begin{aligned} \min_{w \in \mathbb{R}^D} \quad & \|w\|^D + C\sum_{i=1}^{n} \xi_i \\ \text{subject to,} \quad & y_i(x_i^T w) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \dots, n. \end{aligned}$$

which can be shown to be equivalent to that in Equation (3.1) for $C = \frac{1}{\lambda n}$ (and clearly ℓ being the hinge loss). The derivation of a computational procedure to solve the SVM minimization problem requires notions from convex optimization, beyond the scope of this brief introduction. Indeed, it can be shown that the solution of the SVM problem is of the form

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

where the coefficients $\alpha_1, \dots, \alpha_n$ are given by the solution of the following quadratic programming problem

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, \quad \text{subject to,} \quad \alpha_i \geq 0, \quad i = 1, \dots, n. \tag{3.2}$$

The above problem is usually called the dual problem. An interesting feature of the SVM solution is that its solution requires estimating n, rather than D, coefficients and that vector of coefficients can be *sparse*, that is some of its entries can be zero. The input points for which the corresponding coefficients are non zero are called *support vectors*.

3.7 Dealing with an Offset

When considering linear models, especially in relatively low dimensional spaces, it is interesting to consider an offset, that is $f_{w,b}(x) = w^T x + b$. 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 $f_{w,b}$ and still regularize considering only $\|w\|^2$, so that the offset is not penalized. Note that the regularized problem becomes

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

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

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

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

For the SVM algorithm the effect of considering an offset term is also simple, since we simply have to add the constraint

$$\sum_{i=1}^{n} y_i \alpha_i x_i = 0$$

to the dual problems (3.2).