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

Lecture 16- From the Perceptron to SVM

Lecturer: Lorenzo Rosasco

We next introduce the support vector machine discussing one of the most classical learning algorithm, namely the perceptron algorithm.

16.1 Perceptron

The perceptron algorithm finds a linear classification rule according to the following iterative procedure. Set $w_0 = 0$ and update

$$w_i = w_{i-1} + \gamma y_i x_i, \text{ if } \; y_i w^T x_i \leq 0$$

and let $w_i = w_{i-1}$ otherwise. In words, if an example is correctly classified, then the perceptron does not do anything. If the perceptron incorrectly classifies a training example, each of the input weights is moved a little bit in the correct direction for that training example. The above procedure can be seen as the stochastic (sub) gradient associated to the objective function

$$\sum_{i=1}^{n} |-y_i w^T x_i|_+$$

where the $|a|_+ = \max\{0, a\}$. Indeed if $y_i w^T x_i < 0$, then $|-y_i w^T x_i|_+ = -y_i w^T x_i$ and $\nabla|-y_i w^T x_i|_+ = -y_i x_i$, if $y_i w^T x_i > 0$, then $|-y_i w^T x_i|_+ = 0$ hence $\nabla|-y_i w^T x_i|_+ = 0$. Clearly an off-set can also be considered, replacing $w^T x$ by $w^T x + b$ and analogous iteration can ve derived.

The above method can be shown to converge for $\gamma = const.$ if the data are linearly separable. If the data are not separable with a constant step size the perception will typically cycle. Moreover, the perceptron does not implement any specific form of regularization so in general it is prone to overfit the data.

16.2 Margin

The quantity $\alpha = y w^T x$ defining the objective function of the perceptron is a natural error measure and is sometimes called the *functional margin*. Next we look at a geometric interpretation of the functional margin that will lead to a different derivation of Tikhonov regularization for the so called hinge loss function. We begin considering a binary classification problem where the classes are linearly separable.

Consider the decision surface $D = \{x : w^T x = 0\}$ defined by a vector w and x such that $w^T x > 0$. It is easy to check that, the projection of x on D is a point x_w satisfying,

$$x_w = x - \beta \frac{w}{\|w\|}$$

where β is the distance between x and D. Clearly $x_w \in D$ so that

$$w^T(x - \beta\frac{w}{\|w\|}) = 0 \Leftrightarrow \beta = \frac{w^T}{\|w\|}x.$$

If x is x such that $w^T x < 0$ then $\beta = -\frac{w^T}{\|w\|}x$, so that generally we have

$$\beta = y\frac{w^T}{\|w\|}x$$

The above quantity is often called the geometric margin and clearly if $\|w\| = 1$ is coincides with the geometric margin. Note that the margin is scale invariant, in the sense that $\beta = y\frac{w^T}{\|w\|}x = y\frac{2w^T}{\|2w\|}x$, as is the decision rule $\text{sign}(w^T x)$.

16.3 Maximizing the Margin

Maximizing the margin is a natural approach to select a linear separating rule in the separable case. More precisely consider

$$\begin{aligned} \beta_w &= \min_{i=1,\ldots,n} \beta_i, \quad \beta_i = y_i\frac{w^T}{\|w\|}x_i, \quad i = 1,\ldots,n, \\ \max_{w\in\mathbb{R}^D} & \quad \beta_w, \quad \text{subj. to,} \quad \beta_w \geq 0, \quad \|w\| = 1. \end{aligned} \tag{16.1}$$

Note that the last constraint is needed to avoid the solution $w = \infty$ (check what happens if you consider a solution w and then scale it by a constant kw).

In the following we manipulate the above expression to obtain a problem of the form

$$\min_{w\in\mathbb{R}^D} F(w), \quad Aw + c \geq 0,$$

where F is convex, A is a matrix and c a vector. These are convex programming problem which can efficiently solved.

We begin by rewriting problem (16.1) by introducing a *dummy* variable $\beta = \beta_w$ to obtain

$$\max_{(w,\beta)\in\mathbb{R}^{D+1}} \beta, \quad \text{subj. to,} \quad y_i\frac{w^T}{\|w\|}x_i \geq \beta; \beta \geq 0, \|w\| = 1$$

(we are basically using the definition of minimum as the maximum of the infimal points). We next would like to avoid the constraint $\|w\| = 1$. It can be shown that the above problem is equivalent to considering

$$\max_{(w,\alpha)\in\mathbb{R}^{D+1}} \frac{\alpha}{\|w\|}, \quad \text{subj. to,} \quad y_i w^T x_i \geq \alpha; \alpha \geq 0.$$

with $\beta = \frac{\alpha}{\|w\|}$, where the key idea is that the latter problem is scale invariant. More precisely that we can always restrict yourselves to $\|w\| = 1$ by appropriately rescaling the solutions.

Using again scale invariance (check what happens if you consider a solution w and then scale it by a constant $(kw, k\alpha)$), without loss of generality we can fix $\alpha = 1$ to obtain

$$\max_{w\in\mathbb{R}^D} \frac{1}{\|w\|}, \quad \text{subj. to,} \quad y_i w^T x_i \geq 1 \quad , i = 1, \ldots, n,$$

or equivalently

$$\min_{w\in\mathbb{R}^D} \frac{1}{2}\|w\|^2, \quad \text{subj. to,} \quad y_i w^T x_i \geq 1 \quad , i = 1, \ldots, n, \tag{16.2}$$

In the above reasoning we assumed data to be separable if this is not the case one could considering *slack* variables $\xi = (\xi_1, \ldots, \xi_n)$ to relax the constraints in the above problem, considering

$$\min_{w\in\mathbb{R}^D, \xi\in\mathbb{R}^n} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}, \quad \text{subj. to}, y_i w^T x_i \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad , i = 1, \ldots, n. \tag{16.3}$$

16.4 From Max Margin to Tikhonov Regularization

Note that $\xi_i = \max\{0, 1 - y_i w^T x_i\} = |1 - y_i w^T x_i|_+$, for all $i = 1, \ldots, n$. Then if we set $\lambda = \frac{1}{2Cn}$, we have that problem (16.3) is equivalent to

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

16.5 Computations

The derivation of a solution to the SVM problem requires notions of convex optimization, specifically considering so called Lagrangian duality. Indeed, it can be shown that the solution of problem (16.3) is of the form

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

where the coefficients α^i for $i = 1, \ldots, n$ are given by the solution of the so called dual problem,

$$\min_{\alpha\in\mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, \quad \text{subject to,} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n, \tag{16.4}$$

where in particular it can be shown that

$$\alpha_i = 0 \Longrightarrow y_i w^T x_i \geq 1.$$

16.6 Dealing with an off-set

Finally, it can be shown that the above reasoning can be generalized to consider an offset, that is $w^T x + b$, in which case we simply have to add the constraint

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

to the dual problem (16.4).