29-Jan-01

Lecture-5

Supplementary

Aim Using the key theorem (Thm. 4.7 in [Bartlett and Shawe-Taylor 1999]) we wish to motivate the (soft-margin) SVM formulation.

Working By the ~~data~~ VC-type inequality of Thm. 4.7, we need to

minimize $\boxed{\sum_{i=1}^{m} 1_{\{y_i(w^T x_i - \theta) < \gamma\}}}$ and $\boxed{\text{maximize}}$ $\boxed{\gamma}$ while being in the assumptions of the theorem.

margin

Empirical risk

min. the Φ term

~~i.e. $\min_{w,\theta} \sum_{i=1}^{m} 1_{\{y_i(w^T x_i - \theta) < \gamma\}}$~~

~~s.t. $\|w\| \leq 1$~~

One way of achieving this is by using principle of (data dependent) SRM:

(i) First we order the functions in $\mathcal{F}$ according to decreasing γ i.e. increasing complexity.

(ii) For ~~this~~ each given subset of functions in this hierarchy, minimize the empirical risk.

~~(iii) Select ~~the~~ subset of fu~~

(iii) After considering the many subsets, we can choose the best may be based on the VC-bound of true risk.

(ii) can be done by solving the following optimization problem.

$$\min_{w,\theta,\gamma} \sum_{i=1}^{m} 1_{\{y_i(w^T x_i - \theta) < \gamma\}}$$

Empirical risk.

$$\text{s.t.} \quad \boxed{\gamma \geq \gamma_0}, \quad \|w\| \leq 1$$

assumption of theorem

γ_0 depends on the subset chosen.

This basically defines the subset of functions.

Now lets re-write above problem in terms of $\bar{w}$ & $\bar{b}$ where

$$\bar{w} = \frac{w}{\gamma}, \quad \bar{b} = \frac{b}{\gamma}.$$

$$\min_{\bar{w}, \bar{b}, \gamma} \sum_{i=1}^{m} 1_{\{y_i(\bar{w}^T x_i - \bar{b}) < 1\}}$$

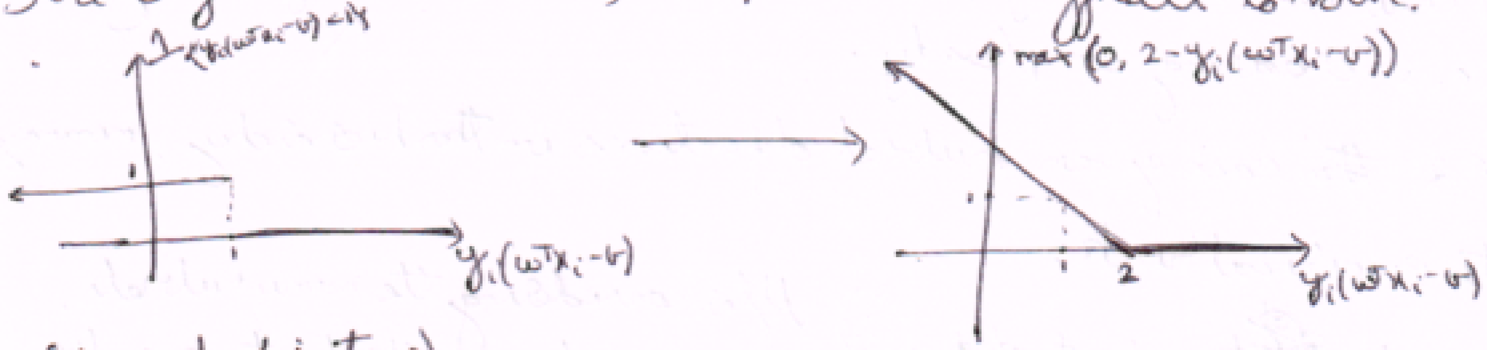
s.t. $\gamma \geq \gamma_0$, $\|\bar{w}\| \leq \frac{1}{\gamma}$

eliminating γ & calling $W_0 = \frac{1}{\gamma_0}$, we have:

$$\min_{w, b} \sum_{i=1}^{m} 1_{\{y_i(w^T x_i - b) < 1\}}$$ → objective is discontinuous & non-convex.

s.t. $\|w\| \leq W_0$ → SOC constraint (second order cone) or Conic quadratic constraint it is convex constraint

Since objective is non-convex, the problem is difficult to solve.



(original objective) (new objective)

So instead of minimizing $1_{x,y}$, we minimize a convex upperbound

We get the following problem:

$$\min_{w, b} \sum_{i=1}^{m} \max(0, 2 - y_i(w^T x_i - b)) \quad \text{s.t. } \|w\| \leq W_0$$

$$\equiv \min_{w, b} \sum_{i=1}^{m} \max(0, 1 - y_i(w^T x_i - b)) \quad \text{s.t. } \|w\| \leq W$$

(where $W = W_0/2$)

Problem I can also equivalently be written as:

$$\text{II} \qquad \min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

regularize

regularization parameter

minimize Empirical risk

This is the final form of the SVM formulation & is an instance of Tikhonov regularization.

Problems I & II are also equivalent to:

$$\text{III} \qquad \min_{w,b,\xi} \ \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \sum_{i=1}^{m}\xi_i \leq E$$

regularize

bound the empirical risk

regularization parameter.

This is an instance of Morozov regularization.

(Unfortunately there is no analytical expression which relates W, C, E)

Also, $\xi_i = \max(0, 1 - y_i(w^T x_i - b))$ is known as hinge-loss function.

One can also employ $\left(\max(0, 1 - y_i(w^T x_i - b))\right)^d$ as loss function.

This is hinge loss function of degree $d \geq 1$.

In fact, the new objective (though is non-differentiable) can be written as:

$$\text{(I)} \quad \min_{w,b,\xi_i} \sum_{i=1}^{m} \xi_i$$

minimize training error

s.t. $\|w\| \le W$, $\xi_i \ge 0$, $\xi_i \ge 1 - y_i(w^T x_i - b)$

"regularize the model"

$$\Downarrow$$

$\xi_i \ge \max(0, 1 - y_i(w^T x_i - b))$

but since obj. is min. ξ_i we will have

$\xi_i = \max(0, 1 - y_i(w^T x_i - b))$ at optimality.

This is an instance of Second Order Cone Program (SOCP) i.e. opt. problem in which objective is linear & constraints are SOCs.

(Note that linear constraints are special cases of SOCs)

Such opt. problems where a model which minimizes empirical fit to data while being "regular" ($\|w\| \le W$) is needed for ~~are known as~~ come under the heading of Ivanov Regularization.

W is called regularization parameter.

In our case, it is actually ~~inver~~ proportional to the VC dim. of the ~~com~~ ~~com~~ classifier.