Regularization and Structual Risk Minimization

Ruixin Guo

September 5, 2023

1 Introduction

Let (x_i, y_i) be the dataset where $i = 1, 2, ..., m$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Let $f(x;\theta)$ be the machine learning model, where $x \in \mathbb{R}^d$ is the input vector, $\theta \in \mathbb{R}^p$ is the parameter vector. We define the empirical risk function as

$$R^{\text{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i;\theta) - y_i)^2$$

Instead of minimizing $R^{\text{emp}}(f)$, we are interested in the following constraint problem

$$\min_{\theta} R^{\text{emp}}(f(x;\theta)) \qquad \text{s.t.} \quad \|\theta\|^2 \leq c \tag{1}$$

where c is a positive constant, and $\|\cdot\|$ is Euclidean norm.

Suppose all the f forms a hypothesis space $\mathcal{F}$, and all the θ forms a parameter space Θ. Each θ determines an f uniquely, so there is a bijection mapping between $\mathcal{F}$ and Θ.

Since $\|\theta\|^2 \leq c$, smaller c means smaller hypothesis space $\mathcal{F}$. The Structural Risk Minimization takes different constants $c_1 > c_2 > ... > c_n$ which yields a series of nested hypothesis spaces $\mathcal{F}_1 \supset \mathcal{F}_2 \supset ... \supset \mathcal{F}_n$. Smaller hypothesis gives better generalization. and by shrinking the size of hypothesis space, we can find a suitable size that enable us to find a function that is close to the minimizer of the true risk.

2 The Relationship between Regularization and Structural Risk Minimization

Let $g(\theta) = R^{\text{emp}}(f(x;\theta))$. We can write equation (1) as the following Lagrange function

$$L(\theta, \lambda) = g(\theta) + \lambda(\|\theta\|^2 - c) \tag{2}$$

This Lagrange function is of inequality constraint since $\|\theta\|^2 \leq c$. We want to find the θ to minimize $L(\theta, \lambda)$. The solution has two cases:

(1) $\lambda = 0$ when $\|\theta\|^2 < c$. This means when the minimizer θ^* is inside the boundary $\|\theta\|^2 = c$, the constraint does not take effect.

(2) $\lambda > 0$ when $\|\theta\|^2 = c$. This means θ^* is on the boundary and the constraint takes effect.

The system of equations that unites the above two cases to solve the Lagrange function is known as the KKT conditions [1], which are as follows:

$$\frac{\partial L}{\partial \theta} = \frac{\partial g}{\partial \theta} + 2\lambda\theta = 0 \tag{3}$$

$$\frac{\partial L}{\partial \lambda} = \|\theta\|^2 - c \leq 0 \tag{4}$$

$$\lambda \geq 0 \tag{5}$$

$$\lambda(\|\theta\|^2 - c) = 0 \tag{6}$$

[1] https://en.wikipedia.org/wiki/Karush-Kuhn-Tucker_conditions

Generally, we get a set of (λ, θ) pairs by equations (3) and (6) as candidate solutions, and check if these pairs satisfy inequality (4) and (5).

By equations (3) and (6), given c, we can solve θ and λ; given λ, we can solve θ and c. Structural Risk Minimization solves θ by fixing c. Regularization solves θ by fixing λ.

In Regularization, suppose $\lambda > 0$, and (3) and (6) give a set of solution pairs $\{(\theta_j, c_j)\}_{j=1}^n$. Let $c = \max_j\{c_j\}$, then c is independent of θ_j. We can consider c as a function of λ. Hence, (2) can be written as

$$L(\theta, \lambda) = g(\theta) + \lambda(\|\theta\|^2 - c(\lambda)) \tag{7}$$

When λ is fixed, $c(\lambda)$ is a constant. So let

$$L_1(\theta, \lambda) = g(\theta) + \lambda\|\theta\|^2 \tag{8}$$

we have

$$\operatorname*{argmin}_\theta L(\theta, \lambda) = \operatorname*{argmin}_\theta L_1(\theta, \lambda)$$

This means we can solve L_1 instead of L to find θ^*.

3 The Relationship between λ and c

(1) If $c \to \infty$, then $\lambda \to 0$. In Eq (6), when $\|\theta\|^2$ is finite and $c = \infty$, we must have $\lambda = 0$.

(2) If $c \to 0$, then $\lambda \to \infty$. $c = 0$ means $\theta = 0$. By Eq (3), $\lambda = -\frac{\partial g}{\partial \theta}\frac{1}{\theta}$. And by Eq (5), $\lambda \geq 0$. Suppose $-\frac{\partial g}{\partial \theta} \neq 0$ when $\theta = 0$, if λ satisfies the KKT condition, then it must have $\lambda = \infty$.

Therefore, we can consider λ as a function of c. The function has a decreasing trend but may not necessary be non-increasing.

Theorem: Let λ be a continuous function of c and $\lambda \to 0$ when $c \to \infty$. Let $\lambda_i = \lambda(c_i)$ for $i = 1, 2, ..., n$. Then there exists $c_1 < c_2 < ... < c_n$ to make $\lambda_1 > \lambda_2 > ... > \lambda_n$.

Proof: Since $\lambda \to 0$ when $c \to \infty$, the definition of convergence says $\forall \epsilon > 0$, $\exists \delta$ such that $\forall c > \delta, \lambda(c) < \epsilon$.

Let $\epsilon = \lambda_i$, there must exist c_i such that for any $c > c_i$, $\lambda(c) < \lambda_i$.

Let $\lambda_{i+1} < \lambda_i$, there must exist c_{i+1} such that for any $c > c_{i+1}$, $\lambda(c) < \lambda_{i+1}$.

Let $C_1 = \{c : \lambda(c) < \lambda_i\}$, $C_2 = \{c : \lambda(c) < \lambda_{i+1}\}$, we know that $C_2 \subset C_1$. Thus $c_{i+1} > c_i$.