Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Juncai He, Qingguo Hong, Li Jiang.....

Contents

0.1 Binary LR and SVM and their relations 3
0.1.1 Binary SVM 3
0.1.2 Binary Logistic Regression 6

0.1 Binary LR and SVM and their relations

Given a binary linealy separable classification dataset $(x_i, y_i)_{i=1}^N$, where $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$. We use A_1, A_2 to denote the data with label $+1, -1$ respectively. Our goal is to find a $\theta = (w, b)$ where $w \in \mathbb{R}^{1\times d}, b \in \mathbb{R}$ such that the hyperplane $H_\theta = \{x : wx + b = 0\}$ can separate A_1, A_2.

0.1.1 Binary SVM

Binary SVM wants to find the classifiable hyperplane which has the biggest distance with A_1 and A_2.

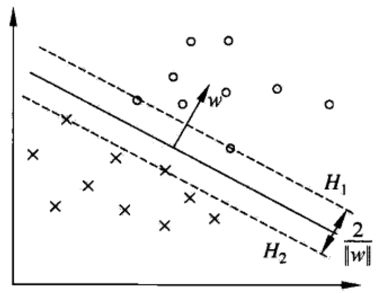
$$\max_{w,b} \frac{\min_i y_i(wx_i + b)}{\|w\|_2} \tag{0.1}$$

Intuitively, the best separating hyperplane H are only determined by those data points who are closest to H. Those data points are called support vector, and this method are called support vector machine.

Without loss of generality, we may restrict the norm of $\|w\|$ to be 1, which leads to a equivalent optimization problem

$$\max_{\|w\|_2=1} \min_i y_i(wx_i + b) \tag{0.2}$$

Actually, we can prove $\text{argmax}_{\|w\|_2=1} \min_i y_i(wx_i + b)$ is nonempty, but here we just admit this fact and only prove the uniqueness of the solution.



Lemma 1. *If A_1, A_2 are linearly separable, then*

$$\operatorname*{argmax}_{\|w\|_2=1} \min_i y_i(wx_i + b) \tag{0.3}$$

is a singleton set.

Proof. Denote $m(w, b) = \min_i y_i(wx_i + b)$. Notice that $m(w, b)$ is a concave homogeneous function w.r.t w, b and $\|\cdot\|_2$ is a strictly convex norm. Suppose there are two solution (w_1, b_1) and (w_2, b_2) such that $w_1 \neq w_2$, take $\overline{w} = \frac{w_1+w_2}{2}, \overline{b} = \frac{b_1+b_2}{2}$, we must have

$$m(\overline{w}, \overline{b}) \geq \frac{m(w_1, b_1) + m(w_2, b_2)}{2} = \max_{\|w\|_2=1} m(w, b), \tag{0.4}$$

and

$$\|\overline{w}\|_2 < 1. \tag{0.5}$$

So

$$m(\frac{\overline{w}}{\|\overline{w}\|_2}, \frac{\overline{b}}{\|\overline{w}\|_2}) = \frac{m(\overline{w}, \overline{b})}{\|\overline{w}\|_2} > \max_{\|w\|_2=1} m(w, b), \tag{0.6}$$

which leads to a contradiction. So all the solution must have the same w, we denote it as w^*. Then if (w^*, b^*) is a solution of problem (0.3), we must have

$$b^* \in \operatorname*{argmax}_b m(w^*, b) \tag{0.7}$$

Actually,

$$m(w^*, b) = \min\{b + \min_{x \in A_1} w^* x, -b + \min_{x \in A_2}(-w^* x)\}, \tag{0.8}$$

easy to observe that $\operatorname{argmax}_b m(w^*, b)$ is a singleton set and

$$b^* = \frac{\min_{x \in A_2}(-w^* x) - \min_{x \in A_1} w^* x}{2}. \tag{0.9}$$

□

Denote

$$\theta^*_{SVM} = (w^*_{SVM}, b^*_{SVM}) = \operatorname*{argmax}_{\|w\|=1} \min_i y_i(wx_i + b). \tag{0.10}$$

Theorem 1. *w^*_{SVM} must be a linear combination of $x_i^T, i = 1, 2, \cdots, N$.*

Proof. Denote

$$S = \text{span}\{x_i^T\}_{i=1}^N \tag{0.11}$$

Then we have

$$\mathbb{R}^{1\times d} = S \oplus^{\perp} S^{\perp} \tag{0.12}$$

So w^*_{SVM} can be uniquely decomposed as $w^*_{SVM} = w^*_S + w^*_{S^\perp}$ where $w_S \in S$ and $w^*_{S^\perp} \in S^\perp$. We will prove that $w^*_{S^\perp} = 0$. Suppose not, we have

$$\|w^*_S\|_2 < \|w^*\|_2 = 1. \tag{0.13}$$

Notice that

$$w^*_{SVM} x_i = w^*_S x_i, \ \forall i = 1, 2, \cdots, N. \tag{0.14}$$

Thus we have

$$\min_i y_i(w^*_{SVM} x_i + b^*) = \min_i y_i(w^*_S x_i + b^*) \tag{0.15}$$

So

$$\min_i y_i(w^*_{SVM} x_i + b^*_{SVM}) < \frac{\min_i y_i(w^*_S x_i + b^*_{SVM})}{\|w^*_S\|} = \min_i y_i(\frac{w^*_S}{\|w^*_S\|_2} x_i + \frac{b^*_{SVM}}{w^*_S}), \tag{0.16}$$

which leads to a contradiction to the definition of θ^*_{SVM}. □

We may rewrite the SVM problem as

$$\min_{w,b} \|w\|^2, \tag{0.17}$$

$$s.t.\ y_i(wx_i + b) \geq 1, \ \forall i. \tag{0.18}$$

We can simply prove that the solution of (0.20) is θ^*_{SVM} multiplies a positive scalar. So it still satisfies the representer theorem. Thus we can restrict w to be in the set S. Assume that

$$w = \sum_{i=1}^N \alpha_i x_i^T, \tag{0.19}$$

Denote $\alpha = (\alpha_1, \cdots, \alpha_N)^T$, and $D \in \mathbb{R}^{N\times N}$ where $D_{ij} =< x_i, x_j >$. We can rewrite the problem (0.20) as

$$\min_{w,b} \alpha^T D \alpha, \tag{0.20}$$

$$s.t.\ y_i(\sum_{j=1}^N < x_j, x_i > \alpha_j + b) \geq 1, \ \forall i. \tag{0.21}$$

We can see that the whole problem is only determined by the inner product of data points but not the data itself. What we called kernel method is just use a symmetric positive definite kernel function to replace the inner product. Such kernel function can be regarded as a inner product of some feature space.

0.1.2 Binary Logistic Regression

For binary logistic regression, our score mapping can be written as $\begin{pmatrix} \frac{1}{1+e^{-(wx+b)}} \\ \frac{1}{1+e^{wx+b}} \end{pmatrix}$. We can observe that, (w, b) is classifiable if and only if

$$\frac{1}{1+e^{-y_i(wx+b)}} > \frac{1}{2},\ \forall i = 1, 2 \cdots, N. \tag{0.22}$$

So we may consider to maximize following objetive

$$P(\theta) = \prod_{i=1}^{N} \frac{1}{1+e^{-y_i(wx+b)}}, \tag{0.23}$$

which is equivalent to minimize

$$L(\theta) = -\log P(\theta) = \sum_{i=1}^{N} -\log(1+e^{-y_i(wx+b)}), \tag{0.24}$$

Lemma 2. *$L(\theta)$ is a strictly convex function without any global minima.*

To let the above problem have a global minima, we may add a L_2 regularization term as following

$$\mathcal{L}(\theta, \lambda) = L(\theta) + \lambda \|w\|_2^2 = \sum_{i=1}^{N} -\log(1+e^{-y_i(wx+b)}) + \lambda \|w\|_2^2, \tag{0.25}$$

Actually, we can prove $\operatorname{argmin}_{w,b} L(\theta, \lambda)$ is nonempty for λ sufficiently small, but here we just admit this fact and only prove the uniqueness of the solution.

Lemma 3. *If A_1, A_2 are linearly separable, then*

$$\operatorname*{argmin}_{w,b} L(\theta, \lambda) \tag{0.26}$$

is a singleton set for λ sufficiently small.

Proof. Because $L(\theta)$ is strictly convex w.r.t. θ and $\|w\|^2$ is convex w.r.t. θ, so $\mathcal{L}(\theta, \lambda) = L(\theta) + \lambda \|w\|_2^2$ is stricly convex w.r.t. θ, which implies our result directly. □

For λ sufficiently small, denote

$$\theta_{LR}(\lambda) = (w_{LR}(\lambda), b_{LR}(\lambda)) = \operatorname*{argmin}_{w,b} L(\theta, \lambda). \tag{0.27}$$

Theorem 2. *If A_1, A_2 are linearly separable, then $\frac{\theta_{LR}(\lambda)}{\|w_{LR}(\lambda)\|}$ converge to θ^*_{SVM} as $\lambda \to 0$, i.e.*

$$\theta^*_{SVM} = \lim_{\lambda \to 0} \frac{\theta_{LR}(\lambda)}{\|w_{LR}(\lambda)\|}. \tag{0.28}$$