Learning with Sparsity

Zhenwei Lin

日期：2020 年 8 月 15 日

摘　　要

本文主要为从优化角度来对监督学习进行介绍

关键词： loss function; penalty function

1 Loss function

1.1 linear regression

$$
\begin{aligned}
l(y, f(x)) =& (y - f(x))^2 \\
H(f) =& \left\{ f | f(x) = a^T x + b \right\} \\
\Rightarrow \min_l =& \min \frac{1}{N} \sum_{i=1}^{N} (y_i - a_i^T x_i - b)^2
\end{aligned}
$$

1.2 LAD regression

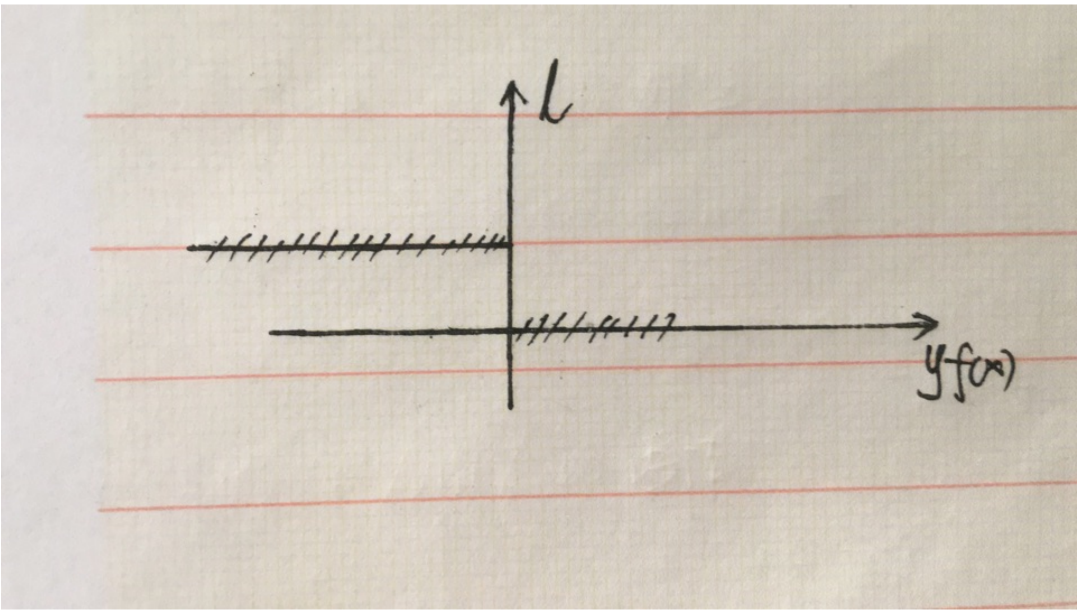
$$
l(y, f(x)) = |y - f(x)|
$$

1.3 0-1 loss-function

$$
l(y, f(x)) = \begin{cases} 1 & y = f(x) \\ 0 & y \neq f(x) \end{cases}
$$

1.4 sign function

$$
\begin{aligned}
f : X \rightarrow & Y \subseteq R \\
G(x) =& Sign(f(x)) \\
l(y, G(x)) =& \begin{cases} 1 & y \cdot f(x) < 0 \\ 0 & y \cdot f(x) \geq 0 \end{cases}
\end{aligned}
$$



为了逼近上面这个函数曲线，我们在不同模型中选取了不同的函数来逼近，具体的有 logistic regression, Adaboost,SVM.

1.5 logistic regression

$$l(y, f(x)) = log(1 + exp(yf(x)))$$

1.6 Adaboost

$$l(y, f(x)) = exp(yf(x))$$

1.7 SVM

$$l(y, f(x)) = (1 + yf(x))_+$$

2 logistic regression

$$
\begin{aligned}
D = \{(x_i, y_i)\}_{i=1}^{N}, y_i \in \{-1, 1\} \\
p(Y_i = 1) = \frac{exp(x"_i^T \beta)}{1 + exp(x_i^T \beta)} = \frac{1}{1 + exp(-x_i^T \beta)} \\
p(Y_i = -1) = \frac{1}{1 + exp(x_i^T \beta)} \\
p(Y_i) = \frac{1}{1 + exp(-y_i x_i^T \beta)} \\
\Rightarrow \min_{\beta} \sum_{i=1}^{N} log(1 + exp(-y_i x_i^T \beta))
\end{aligned}
\tag{1}
$$

如何解这个最小化问题呢？

$$
\begin{aligned}
f(z) =& log(1 + exp(-z)) \\
f'(z) =& -\frac{exp(-z)}{1 + exp(-z)} \\
f''(z) =& f(z)(1 - f(z)) \leq \frac{1}{4} \\
\nabla l(\beta) =& -\sum_{i=1}^{N} \frac{exp(y_i x_i^T \beta)}{1 + exp(y_i x_i^T \beta)} \cdot y_i x_i = X^T diag(Y) P \quad P = \begin{bmatrix} \frac{exp(-y_1 x_1^T \beta)}{1+exp(-y_1 x_1^T \beta)} \\ \vdots \end{bmatrix} \\
\nabla^2 l(\beta) =& \sum_{i=1}^{N} \frac{exp(y_i x_i^T \beta)}{1 + exp(y_i x_i^T \beta)} y_i^2 x_i x_i^T = X^T diag(P(1-P)) X \leq \frac{1}{4} \lambda_m ax(X^T X)
\end{aligned}
$$

因此可以采用 $\frac{1}{\beta} = \frac{4}{\lambda_{max}(X^T X)}$ 的 step 来对其进行下降

2.1 newton's method

$$
\begin{aligned}
\beta_{t+1} =& \beta_t + (\nabla^2 l(\beta_t))^{-1} \nabla l(\beta_t) \\
=& \beta_t + (X^T w X)^{-1} X^T diag(Y) P \quad denote w = diag(P(1-P)) \\
=& (X^T w X)^{-1} X^T w \underbrace{(X\beta_t + X^T diag(Y) P)}_{z_t}
\end{aligned}
$$

这个其实对应的是加权最小二乘的解 $argmin_{\beta} \left\{ (z_t - X\beta)^T w (z_t - X\beta) \right\}$

若其中同样存在不可逆的问题，则选择 sparse logistic regression 的方法来建模

$$
\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} log(1 + exp(-y_i x_i^T \beta)) + \lambda_N \|\beta\|_1 \right\}
$$

3 AdaBoost

Q:Can we construct a strong classifier based on existed weak classifiers?

Algorithm 1: AdaBoost

Input $\{(x_i, y_i)\}_{i=1}^{N}$, $y_i \in \{-1, 1\}$, $m = 1$,sample weight $w_i = \frac{1}{N}$;

while $m \leq M$ **do**

step1:Fit a classifier $G_m(x)$ based on train set D

$$G_m(x) = argmin \sum_{i=1}^{N} I(y_i \neq G_m(x_i)) \cdot w_i^m$$

step2: 计算错分率

$$err_m = \frac{\sum_{i=1}^{N} w_i^m I(y_i \neq G_m(x_i))}{\sum w_i^m}$$

step3: 更新权重, 分对权重不变，分错权重增加

$$w_i^{m+1} \leftarrow w_i^m exp(\alpha_m \cdot I(y_i \neq G_m(x_i)))$$

$$\alpha_m = log(\frac{1 - err_m}{err_m})$$

$$m \leftarrow m + 1$$

step4:renormalization

$$\sum w_i^{m+1} = 1$$

end

Output: Majority Volting

$$G(x) = Sign\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

总共有 M 个弱分类器

算法 2，之后我们再说明二者是等价的。

Algorithm 2: Forward Stepwise Additive Model

Input $\{(x_i, y_i)\}_{i=1}^{N}$, $y_i \in \{-1, 1\}$, $f^{(0)}(x_i) = 0, m = 1$;

while $m \leq M$ **do**

step1:Compute

$$(\hat{\beta}_m, G_m(x)) = argmin \sum l(y_i, f^{(m)}(x_i) + \beta_m G_m(x))$$

step2:

$$f^{(m+1)}(x) = f^{(m)}(x) + \hat{\beta}_m G_m(x)$$

$$m \leftarrow m + 1$$

end

Output:

$$G(x) = Sign(f^{(M)}(x)) = Sign\left\{\sum_{m=1}^{M} \hat{\beta}_m G_m(x)\right\}$$

定理 3.1. AdaBoost is equivalent to the forward Stepwise Additive Model by using the exponential error $l(z) = exp(-z)$

证明.

$$
\begin{aligned}
\sum_{i=1}^{N} l(y_i, f^m(x_i) + \beta_m G_m(x_i)) &= \sum_{i=1}^{N} exp\left\{-y_i \cdot (f^{(m)}(x_i) + \beta_m G_m(x_i))\right\} \\
&= \sum_{i=1}^{N} exp(-y_i f^{(m)}(x_i)) exp(-y_i \beta_m G_m(x_i)) \\
&= \sum_{i=1}^{N} w_i^{(m)} exp(-\beta_m y_i G_m(x_i)) \\
&= e^{\beta_m} \sum_{y_i \neq G_m(x_i)} w_i^{(m)} + e^{-\beta_m} \sum_{y_i = G_m(x_i)} w_i^{(m)} \\
&= e^{\beta_m} \sum_{y_i \neq G_m(x_i)} w_i^{(m)} - e^{-\beta_m} \sum_{y_i \neq G_m(x_i)} w_i^{(m)} + e^{-\beta_m} \sum w_i^{(m)}
\end{aligned}
$$

$$(e^{\beta_m} - e^{-\beta_m}) \sum w_i^m \cdot I(G_m(x_i) \neq y_i) + \sum w_i^m e^{-\beta_m} \triangleq l(\beta)$$

$$\Rightarrow G_m(x) = argmin_{G_m(x)} \sum w_i^m I(G(x_i) \neq y_i)$$

$$\nabla l(\beta_m) = (e^{\beta_m} + e^{-\beta_m}) \sum w_i^m I(G_m(x) \neq y_i) - e^{-\beta_m} \sum w_i^m = 0$$

$$\Rightarrow \hat{\beta}_m = \frac{1}{2} log \frac{1 - err_m}{err_m} = \frac{\alpha_m}{2}$$

$$G(x) = Sign(\sum \hat{\beta}_m G_m(x)) = Sign(\sum \frac{\alpha_m}{2} G_m(x))$$

二者的优化问题，故二者等价 □