ML Note

qhy

2017 年 7 月 31 日

目录

1 SVM

定义标签:$\hat{y}=\{+1,-1\}$,+1表示正类,-1表示负类定义输出:

$$g(x)=\begin{cases}+1 & ,f(x)>0\\ -1 & ,f(x)>1\end{cases} \tag{1}$$

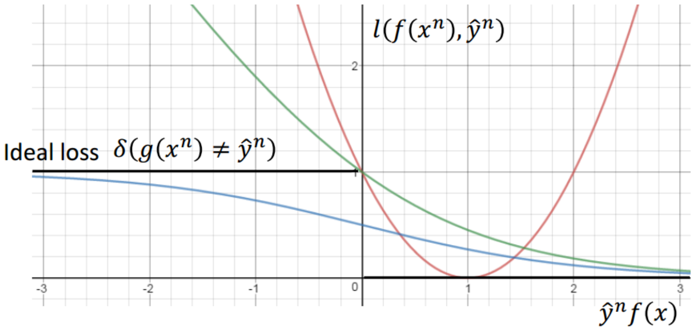
定义loss functon:

$$L(f)=\sum_n I(g(x^n \neq \hat{y}^n)) \tag{2}$$

也就是统计分类错误的个数,从更一般的角度来看,就是每分类错误一个,就给一定的惩罚

分析:$\hat{y^n} f(x^n)>0$的时候,分类正确,无论是正类还是负类,

如果把这个看做一个新的变量t,那么损失函数(一个数据的损失函数)就可以定义成$t>0$时无惩罚,$t<0$的时候给出一定的惩罚。



理想的情况就是图1中的黑线,分类错误就给出惩罚。

由于此loss function使用梯度下降算法麻烦,故使用一个惩罚函数l代替I:

$$L(f)=\sum_n I(f(x^n,\hat{y}^n)) \tag{3}$$

红色线为**Square loss**:

$$l(f(x^n,\hat{y}^n)=(\hat{y}^n f(x^n)-1)^2 \tag{4}$$

显然是不合理的,因为它对于分类正确的结果也给出了惩罚。

蓝色线为**sigmoid+ square loss**:

$$l(f(x^n,\hat{y}^n)=(\sigma(\hat{y}^n f(x^n))-1)^2 \tag{5}$$

这样,得到的结果是合理的,对于分类正确,确定性高的给出低惩罚,对于分类错误,甚至错误的离谱的,则给出大的惩罚

但是有一个问题就是,对于分类错误的结果,它的惩罚很低。

绿色线为**Sigmoid + cross entropy**

$$l(f(x^n,\hat{y}^n)=\ln(1+exp(-y^n f(x^n))) \tag{6}$$

蓝色线为**hinge loss**:

$$l(f(x^n,\hat{y}^n)=max(0,1-y^n f(x^n)) \tag{7}$$

对于分类正确的结果,只要$yf(x)<1$的数据还是会给出惩罚,而SVM采用的就是$HingeLoss$

从$f(x)$来看,$f(x)$为分界线,只要在$f(x)$上下距离为1的样本都会给出惩罚。

1.1 线性SVM

模型如下:

- Step 1: Function (Model)

$$f(x)=\sum_i w_i x_i + b = \underbrace{\boxed{\begin{bmatrix} w \\ b \end{bmatrix}}}_{\text{}}^{\text{New w}} \cdot \underbrace{\boxed{\begin{bmatrix} x \\ 1 \end{bmatrix}}}_{\text{New x}} = w^T x$$

- Step 2: Loss function

convex

$$L(f)=\sum_n l(f(x^n),\hat{y}^n) + \lambda \|w\|_2$$

$$l(f(x^n),\hat{y}^n) = max\big(0, 1-\hat{y}^n f(x)\big)$$

梯度下降:

$$\frac{\partial L(f)}{\partial w_i} = \sum_n \underbrace{-\delta(\hat{y}^n f(x^n) < 1)\hat{y}^n}_{c^n(w)} x_i \qquad w_i \leftarrow w_i - \eta \sum_n c^n(w) x_i^n$$

1.2 线性SVM2

至此,线性SVM结束,但是这个SVM和之前见到的SVM貌似不太一样?对偶问题,SMO,等等这些东东呢?

另一种描述:

Minimizing loss function L:

$$L(f)=\sum_n \varepsilon^n + \lambda \|w\|_2$$

$$\boxed{\varepsilon^n = max\big(0, 1-\hat{y}^n f(x)\big)}$$

ε^n: slack variable

Quadradic programming problem

$$||$$

$$\boxed{\begin{array}{l} \varepsilon^n \geq 0 \\ \varepsilon^n \geq 1-\hat{y}^n f(x) \Rightarrow \hat{y}^n f(x) \geq 1-\varepsilon^n \end{array}}$$

经过这样处理,原问题从最小化:

$$L(f)=\sum_n l(f(x^n),\hat{y}^n) + \lambda \|w\|_2$$

变成了最小化:

$$L(f)=\sum_n \varepsilon^n + \lambda \|w\|_2 \quad \text{s.t.} \quad \hat{y}^n f(x) \geq 1-\varepsilon^n$$

进一步处理,问题变成最小化:

$$L(f)= \lambda \|w\|_2 \quad \text{s.t.} \quad \hat{y}^n f(x) \geq 1$$