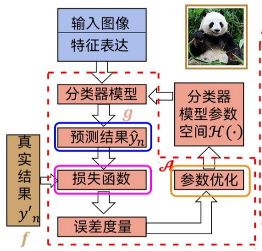
25.1 Logistic 回归问题

Lin=0 时收敛

无法反映两个同loss分类器的优劣(泛化能力)

未在分类感知器中：输出空间离散，损失函数离散，逻辑斯谛实现软分类



逻辑斯蒂回归：软分类

$f(\mathbf{x}_n) = y'_n = P(+1|\mathbf{x}_n) \in [0,1]$

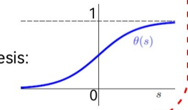
$\hat{y}_n = sign(P(+1|\mathbf{x}_n) - 0.5) \in \{1,-1\}$ $\Rightarrow \hat{y} = h(x) = \theta(w^T x) = \frac{1}{1+\exp(-w^T x)}$

$\mathbf{x} = (x_1, x_2, \dots x_d)^T$

$s = \sum_{i=0}^{d} w_i x_i$

logistic hypothesis:

$h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$



※ 逻辑斯谛函数 $\theta(-\infty)=0$；$\theta(0)=\frac{1}{2}$；$\theta(\infty)=1$ 平滑而单调

$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$　※ $\frac{d\theta(s)}{ds} = \theta(s)(1-\theta(s))$

$h(x) = \theta(w^T x) = \frac{1}{1+\exp(-w^T x)}$

25.2 Logistic 回归损失函数

感知器(线性分类)：$\hat{y}_{n(t)} = sign(w_t^T x_{n(t)})$，$L_{in} = \sum_{n=1}^{N} \mathbb{I}(y_n \neq \hat{y}_n)$

线性回归：$\hat{y}_n = w^T x_n$，$L_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2$

若 Logistic 回归使用MSE作损失函数，$\nabla L_{in}(w,x,y) = 2(\theta(yw^Tx)-1)\theta(yw^Tx)(1-\theta(yw^Tx))yx$，且 Label 为离散值

$\frac{\partial\theta(z)}{\partial z}$ $\begin{cases} z>0，接近0 \\ z<0，接近0 \end{cases}$

※ Logistic 回归最优解

$g = \underset{w}{\arg\min} \frac{1}{N}\sum_{n=1}^{N} -\ln\theta(y_n w^T x_n) = \underset{w}{\arg\min} \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n w^T x_n))$

定义交叉熵损失函数 $L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n w^T x_n))$

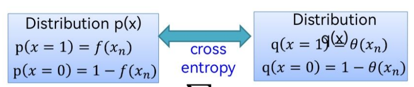
$\ln(1+\exp(-ys))$

令 $s=w^T x$　忽略常数

$L_{in} = \frac{1}{N}\sum_{n=1}^{N}\ln(1+\exp(-ys))$

$\Rightarrow H(f(x_n), \theta(x_n)) = \sum_n [\ln(1+\exp(-ys))]$

交叉熵介绍:



$s = w^T x$

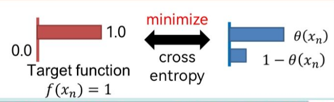
$\theta(s) = \frac{1}{1+\exp(-s)}$

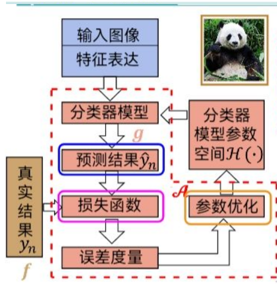
$H(p,q) = -\sum_x p(x)\ln(q(x))$

$H(f(x_n), \theta(x_n)) = \sum_n -[f(x_n)\ln\theta(x_n) + (1-f(x_n))\ln(1-\ln\theta(x_n))]$

$H(f(x_n), \theta(x_n)) = \sum_n -[\ln\theta(x_n)]$

$H(f(x_n), \theta(x_n)) = \sum_n [\ln(1+\exp(-ys))]$





感知器(线性分类)

$\hat{y}_{n(t)} = sign(\mathbf{w}_t^T \mathbf{x}_{n(t)})$

$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$

线性回归

$\hat{y}_n = \mathbf{w}^T \mathbf{x}_n$

$L_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n - y_n)^2$

逻辑斯蒂回归

$\hat{y}_n = h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+exp(-\mathbf{w}^T\mathbf{x})}$

$L_{in} = \sum_{n=1}^{N}\ln(1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n))$

$L_{ce} = -y\ln\hat{y}$

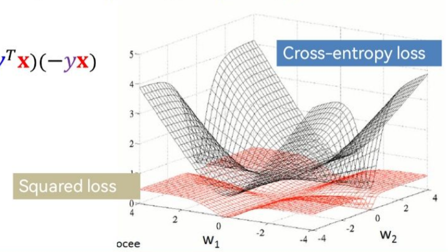
交叉熵损失梯度 $L_{in} = \ln(1+\exp(-y_n \underbrace{w^T x_n}_{s}))$ $\Rightarrow \nabla L_{in}(w,x,y) = \theta(-y_n \overset{s}{\boxed{w^T x_n}})(-yx)$

平方损失的梯度:

$\nabla L_{in}(\mathbf{w}, \mathbf{x}, y) = 2(\theta(y\mathbf{w}^T\mathbf{x}) - 1)\theta(y\mathbf{w}^T\mathbf{x})(1-\theta(y\mathbf{w}^T\mathbf{x}))y\mathbf{x}$

交叉熵损失的梯度:

$\nabla L_{in}(\mathbf{w}, \mathbf{x}, y) = \theta(-y\mathbf{w}^T\mathbf{x})(-y\mathbf{x})$



25.3 Logistic 回归算法

逻辑斯蒂回归

$\hat{y}_n = h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+exp(-\mathbf{w}^T\mathbf{x})}$

$L_{in} = \frac{1}{N}\sum_{n=1}^{N}\ln(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))$

$\nabla L_{in}(\mathbf{w},\mathbf{x},y) = \frac{1}{N}\sum_{n=1}^{N}\theta(-y_n\mathbf{w}^T\mathbf{x}_n)(-y_n\mathbf{x}_n)$

梯度下降法实现逻辑斯蒂回归

- 初始化权向量$\mathbf{w}_0$
- $for\ t = 0,1,2,\dots$ (t 代表迭代次数)　SGD：$\theta(-y_n w_t^T x_n)(-y_n x_n)$

　① 计算梯度：$\nabla L_{in}(\mathbf{w}_t) = \frac{1}{N}\sum_{n=1}^{N}\theta(-y_n\mathbf{w}_t^T\mathbf{x}_n)(-y_n\mathbf{x}_n)$

　② 对权向量$\mathbf{w}_t$进行更新：$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla L_{in}(\mathbf{w}_t)$

…直到$\nabla L_{in}(\mathbf{w}) = \mathbf{0}$，或者迭代足够多次数

返回最终的$\mathbf{w}_{t+1}$作为学到的g

25.4 二元分类线性模型讨论　$s = w^T x$

线性分类(感知器):

$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$

$L_{in} = \sum_{n=1}^{N}[\![y_n \neq \hat{y}_n]\!]$

NP难问题

线性回归:

$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

$L_{in} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n - y_n)^2$

凸函数、易优化、解析解

逻辑斯蒂回归:

$h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$

$L_{in} = \frac{1}{N}\sum_{n=1}^{N}\ln(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))$

凸函数、平滑、梯度下降

线性分类(感知器):

$h(\mathbf{x}) = \text{sign}(s)$

$L_{in} = [\![h(\mathbf{x}) \neq y]\!]$

$L_{0/1}(s,y) = [\![\text{sign}(s) \neq y]\!] = [\![\text{sign}(ys) \neq 1]\!]$

线性回归:

$h(\mathbf{x}) = s$

$L_{in} = (h(\mathbf{x}) - y)^2$

$L_{sqr}(s,y) = (s-y)^2 = (ys-1)^2$

$\nabla L_{in} = (s-y)x$

逻辑斯蒂回归:

$h(\mathbf{x}) = \theta(s)$

$L_{in} = -\ln(h(y\mathbf{x}))$

$L_{ce} = \ln(1+exp(-ys))$

$\nabla L_{in} = \theta(-ys)(-yx_n)$

① 在标签为{+1，-1}的训练样本集 $\mathcal{D}$ 上运行线性回归/逻辑斯蒂回归算法，得到$\mathbf{w}^*$

② 返回分类结果：$g(\mathbf{x}) = sign(\mathbf{w}^{*T}\mathbf{x})$

线性分类(感知器):

优点：样本线性可分时，算法收敛有理论保障

不足：样本非线性可分时NP难问题，可用Pocket算法实现

线性回归:

优点：凸函数，最容易优化，有解析解

不足：当$|ys|$很大时，$L_{0/1}(s,y)$的上界过于宽松

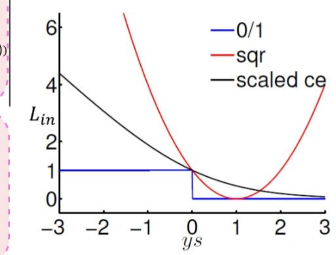
逻辑斯蒂回归:

优点：凸函数，易于优化

不足：当$ys \ll 0$时，$L_{0/1}(s,y)$的上界过于宽松

三种线性模型用于二元分类时(即：$y \in \{+1,-1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\mathbf{w}$ 的内积用 s 表示：$s = \mathbf{w}^T\mathbf{x}$



0/1　$L_{0/1}(s,y) = [\![sign(ys) \neq 1]\!]$

sqr　$L_{sqr}(s,y) = (ys-1)^2$

ce　$L_{ce}(s,y) = ln(1+\exp(-ys))$

Scaled ce　$L_{sce}(s,y) = log_2(1+\exp(-ys))$

$L_{0/1} \leq L_{sqr}$

$L_{0/1} \leq L_{sce}$

$(L_{0/1} \leq L_{ce})$

$\Rightarrow L_{sqr}/L_{ce}$ 很小时 $L_{0/1}$ 也会很小