1. 简述 Logistic Regression

Logistic regression 用来解决二分类问题，

它假设数据服从伯努利分布，即输出为 正 负 两种情况，概率分别为 p 和 1-p，

目标函数 hθ(x; θ) 是对 p 的模拟，p 是个概率，这里用了 p=sigmoid 函数，

所以 目标函数 为：

$$h_\theta(x;\theta) = \frac{1}{1+e^{-\theta^T x}}$$

为什么用 sigmoid 函数？请看：[Logistic regression 为什么用 sigmoid ？]

损失函数是由极大似然得到，

记：

$$\begin{aligned} P(y=1 \mid x;\theta) &= h_\theta(x) \\ P(y=0 \mid x;\theta) &= 1-h_\theta(x) \end{aligned}$$

则可统一写成：

$$p(y \mid x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

写出似然函数：

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

取对数：

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}
$$

求解参数可以用梯度上升：

先求偏导：

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \left(y(1 - g(\theta^T x)) - (1-y) g(\theta^T x) \right) x_j \\
&= \left(y - h_\theta(x) \right) x_j
\end{aligned}
$$

再梯度更新：

$$
\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}
$$

常用的是梯度下降最小化负的似然函数。

2. 先来看常用的几种损失函数：

| 损失函数 | 举例 | 定义 | |

| —— - | — - | —- | —- |

| 0-1损失 | 用于分类，例如感知机 |

$$
L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}
$$

| 预测值和目标值不相等为1，否则为0 |

| 绝对值损失 | |

$$
L(Y, f(X)) = |Y - f(X)|
$$

| |

| 平方损失 | Linear Regression |

$$
L(Y, f(X)) = \sum_N (Y - f(X))^2
$$

| 使得所有点到回归直线的距离和最小 |

| 对数损失 | Logistic Regression |

$$L(Y,P(Y|X)) = -logP(Y|X)$$

| 常用于模型输出为每一类概率的分类器 |

| Hinge损失 | SVM |

$$L(Y,f(X)) = \max(0,1-Yf(x))$$

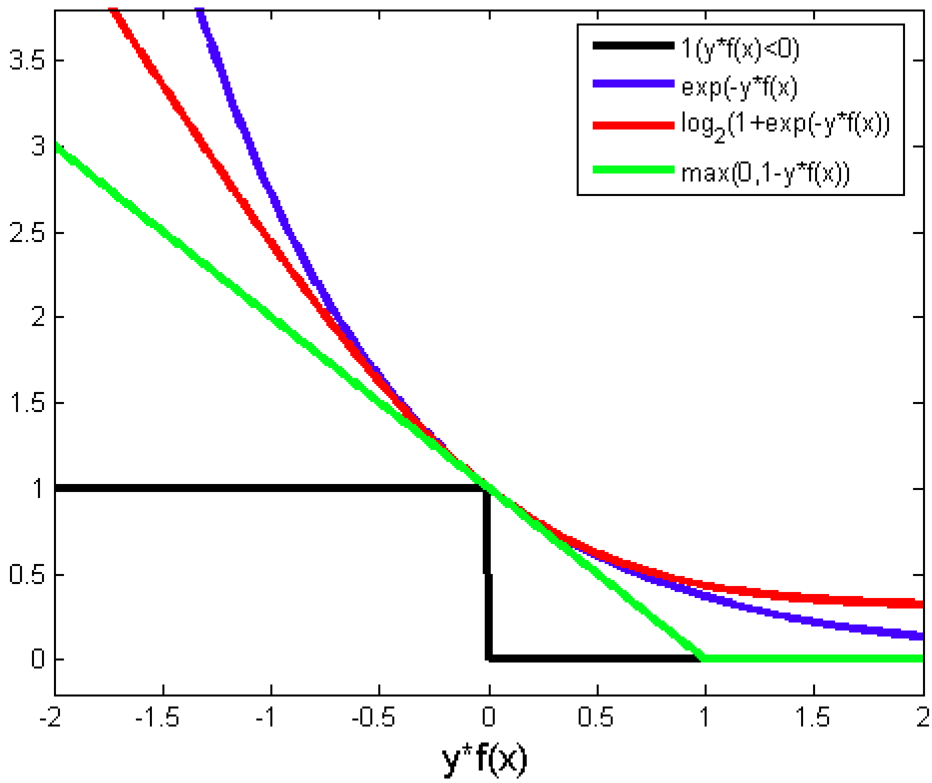
| 用于最大间隔分类 |

| 指数损失 | AdaBoost |

$$L(Y,f(X)) = \exp(-Yf(X))$$

| |

几种损失函数的曲线：



黑色：Gold Stantard

绿色：Hinge Loss中，当 yf(x)>1 时，其损失=0，当 yf(x)<1时，其损失呈线性增长（正好符合svm的需求）

红色 Log、蓝色 Exponential： 在 Hinge的左侧都是凸函数，并且Gold Stantard损失为它们的下界

要求最大似然时(即概率最大化)，使用Log Loss最合适，一般会加上负号，变为求最小

损失函数的凸性及有界很重要，有时需要使用代理函数来满足这两个条件。

3. LR 损失函数为什么用极大似然函数？

1. 因为我们想要让 每一个 样本的预测都要得到最大的概率，

即将所有的样本预测后的概率进行相乘都最大，也就是极大似然函数.

2. 对极大似然函数取对数以后相当于对数损失函数，

由上面 梯度更新 的公式可以看出，

对数损失函数的训练求解参数的速度是比较快的，

而且更新速度只和x，y有关，比较的稳定，

3. 为什么不用平方损失函数

如果使用平方损失函数，梯度更新的速度会和 sigmod 函数的梯度相关，sigmod 函数在定义域内的梯度都不大于0.25，导致训练速度会非常慢。

而且平方损失会导致损失函数是 theta 的非凸函数，不利于求解，因为非凸函数存在很多局部最优解。

什么是极大似然？请看[简述极大似然估计]

学习资料：

https://zhuanlan.zhihu.com/p/25021053

https://www.cnblogs.com/ModifyRong/p/7739955.html https://zhuanlan.zhihu.com/p/34670728

http://www.cnblogs.com/futurehau/p/6707895.html

https://www.cnblogs.com/hejunlin1992/p/8158933.html

http://kubicode.me/2016/04/11/Machine%20Learning/Say-About-Loss-Function/

推荐阅读

历史技术博文链接汇总

也许可以找到你想要的：

[入门问题][TensorFlow][深度学习][强化学习][神经网络][机器学习][自然语言处理][聊天机器人]