机器学习与深度学习面试系列五（逻辑回归）

什么是逻辑回归？与线性回归有什么不同？

逻辑回归处理的是分类问题，线性回归处理的是回归问题，这是两者的最本质的区别。线性回归的一般形式是Y=aX+b，y的取值范围是 $[-\infty, \infty]$ 。对于逻辑回归，就是把Y的结果带入一个非线性变换的**Sigmoid函数（挤压函数）**中，即可得到[0,1]之间取值范围的数S，S可以把它看成是一个概率值P(S = 1|X; w)，它表示当前样本标签为1的概率（1为正样本，0为负样本）。如果我们设置概率阈值为0.5，那么S大于0.5可以看成是正样本，小于0.5看成是负样本，就可以进行分类了。

逻辑回归的一般性公式为： $y=\frac{1}{1+e^{-(w^Tx+b)}}$ ，整理可得： $ln\frac{y}{1-y}=w^Tx+b$ ，这样逻辑回归可以看作是对y的对数几率回归，故称Logistic回归(逻辑回归)。

逻辑回归损失函数是什么？怎么推导？

记 y_i 为样本 x_i 的真实标签， $\hat{y_i}=\frac{1}{1+e^{-(w^Tx_i+b)}}$ 为预测其标签为1的概率，逻辑回归损失函数是交叉熵损失函数： $J=-\sum_{i=1}^{N}(y_ilog\hat{y_i}+(1-y_i)log(1-\hat{y_i}))$ 。

由 $\hat{y_i}=P(y_i=1|x_i)$ ，显然： $1-\hat{y_i}=P(y_i=0|x_i)$ 。将两个式子结合起来： $P(y_i|x_i)=\hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i}$ ，这是一个伯努利分布。对于N个样本，由独立同分布假设可知： $P(Y|X)=\prod_{i=1}^{N}\hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i}$ 。由最大似然法，要估计 $\hat{y_i}$ 中的参数 w 和 b ，只需要求 $P(Y|X)$ 的最大值，即：

$$
\begin{aligned}
w,b &= \arg\max_{w,b} P(Y|X) \\
&= \arg\max_{w,b} logP(Y|X) \\
&= \arg\max_{w,b} log\prod_{i=1}^{N} \hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i} \\
&= \arg\max_{w,b} \sum_{i=1}^{N} log\{\hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i}\} \\
&= \arg\max_{w,b} \sum_{i=1}^{N} \{log\hat{y_i}^{y_i} + log(1-\hat{y_i})^{1-y_i}\} \\
&= \arg\max_{w,b} \sum_{i=1}^{N} \{y_i log\hat{y_i} + (1-y_i)log(1-\hat{y_i})\} \\
&= \arg\min_{w,b} -\sum_{i=1}^{N} \{y_i log\hat{y_i} + (1-y_i)log(1-\hat{y_i})\}
\end{aligned}
$$

可以看出，交叉熵损失函数就是假定 y_i 满足伯努利分布，利用最大似然估计导出的。

逻辑回归的优化过程?

$$
\begin{aligned}
\nabla_w &= \sum_{i=1}^{N} \{y_i \frac{\hat{y_i}(1-\hat{y_i})}{\hat{y_i}} x_i - (1-y_i)\frac{\hat{y_i}(1-\hat{y_i})}{1-\hat{y_i}} x_i\} \\
&= \sum_{i=1}^{N} \{y_i(1-\hat{y_i})x_i - (1-y_i)\hat{y_i}x_i\} \\
&= \sum_{i=1}^{N} \{(y_i-\hat{y_i})x_i\}
\end{aligned}
$$

$$w_t = w_{t-1} - \alpha\nabla_w$$

逻辑回归有什么优点?

1. 对线性关系比较强的拟合效果好
2. 抗噪声能力强
3. 计算快
4. LR能以概率的形式输出结果，而非只是0,1判定，可以做ranking model。

逻辑回归正则化有哪些?

逻辑回归的正则化有L1正则化和L2正则化。

L1正则化：损失函数+ $\alpha||w||_1$ ，其中 $||w||_1$ 为 w 的一范式

L2正则化：损失函数+ $\frac{1}{2}\alpha||w||_2^2$，其中 $||w||_2$ 为 w 的二范式

α 越大正则化强度越大，其中偏置 b 一般不需要正则化。

逻辑回归为什么要对特征进行离散化?

离散化后的特征对异常数据有很强的鲁棒性：比如一个特征是年龄>30是1，否则0。如果特征没有离散化，一个异常数据“年龄300岁”会给模型造成很大的干扰；

特征离散化后，模型会更稳定，比如如果对用户年龄离散化，20-30作为一个区间，不会因为一个用户年龄长了一岁就变成一个完全不同的人。当然处于区间相邻处的样本会刚好相反，所以怎么划分区间是门学问；

特征离散化以后，起到了简化了逻辑回归模型的作用，降低了模型过拟合的风险。

可以参考：

七月在线 七仔：今日面试题分享：逻辑斯特回归为什么要对特征进行离散化
zhuanlan.zhihu.com

如何支持多分类?

可以使用多项逻辑回归(Softmax Regression)来进行分类，假设有k个分类：

$$h_\theta(x_i) = \begin{bmatrix} P(y_i = 1|x_i;\theta) \\ P(y_i = 2|x_i;\theta) \\ \dots \\ P(y_i = k|x_i;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{w_j^T x_i}} \begin{bmatrix} e^{w_1^T x_i} \\ e^{w_2^T x_i} \\ \dots \\ e^{w_k^T x_i} \end{bmatrix}$$

w_1、$w_2 \dots w_k$ 是模型的参数，$\frac{1}{\sum_{j=1}^k e^{w_j^T x_i}}$ 看作归一化参数，对logit做归一化，使所有项的和为1，满足概率分布的要求。当类别k=2时：

$h_\theta(x_i)=\frac{1}{e^{w_1^T x_i}+e^{w_2^T x_i}}\begin{bmatrix} e^{w_1^T x_i} \\ e^{w_2^T x_i} \end{bmatrix}$ ，利用参数冗余的特点，我们对 w_1 、 w_2 同时减去 w_1 ：

$h_\theta(x_i)=\frac{1}{e^{0}+e^{(w_2-w_1)^T x_i}}\begin{bmatrix} e^{0} \\ e^{(w_2-w_1)^T x_i} \end{bmatrix}=\begin{bmatrix} \frac{1}{1+e^{w^T x_i}} \\ 1-\frac{1}{1+e^{w^T x_i}} \end{bmatrix}$ ，该形式与逻辑回归完全一致，因此，多项逻辑回归实际上是二分类逻辑回归在多标签分类下的一种拓展。

当存在样本可能属于多个标签的情况时，我们可以训练k个二分类的逻辑回归 分类器。第i个分类器用以区分每个样本是否可以归为第i类，训练该分类器时，需要把标签重新整理为“第i类标签”与“非第i类标签”两类。通过这样的办法，我们就解决了每个样本可能拥有多个标签的情况。