# 【机器学习算法面试】（一）为什么逻辑回归的损失函数是交叉熵?

原创 潜心 推荐算法的小齿轮 6天前

收录于话题
#机器学习算法面试

2个

## 前言

目前公众号的体裁似乎限定在序列推荐，但这样并不利于广度的学习，因此接下来分享的内容并不会局限于序列推荐（例如上篇文章），会结合目前自己的学习情况，这也是为了扩大读者的范围。当前正在整理机器学习中逻辑回归的基础和面试内容，这里有一个值得思考的问题与大家分享与讨论。

本文约1k字，预计阅读5分钟。

## 概要

逻辑回归（logistic regression）在机器学习中是非常经典的分类方法，周志华教授的《机器学习》书中称其为对数几率回归，因为其属于对数线性模型。

在算法面试中，逻辑回归也经常被问到，常见的面试题包括：

1. 逻辑回归推导；
2. 逻辑回归如何实现多分类？
3. SVM与LR的联系与区别？
4. 逻辑回归反向传播伪代码；

大家可以思考下能不能回答/推导出，但这次讨论的问题是：

> 为什么逻辑回归损失函数是交叉熵？

初看这个问题感觉很奇怪，但是其中的知识包含了LR的推导与理解。在我个人看来，可以从两个角度看待这个问题：

【1】从极大似然估计的角度可以推导出交叉熵；
【2】从KL散度（熵的角度）去理解；

## 极大似然估计

对于逻辑回归，我们一般通过极大似然估计来求解参数$w$。

首先假设两个逻辑回归的两个条件概率：

$$P(Y=1|x)=\frac{exp(w\cdot x)}{1+exp(w\cdot x)}=\pi x$$
$$P(Y=0|x)=\frac{1}{1+exp(w\cdot x)}=1-\pi x$$

学习时，采用极大似然估计来估计模型的参数，似然函数为：

$$L(w)=\prod_{i=1}^{m}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

对数似然函数（采用对数似然函数是因为上述公式的连乘操作易造成下溢）为：

$$L(w)=\sum_{i=1}^{m}y_ilog(\pi(x_i))+(1-y_i)log(1-\pi(x_i))$$

对其求最大值，估计参数$w$:

$$w^*=\underset{w}{argmax}\sum_{i=1}^{m}y_ilog(\pi(x_i))+(1-y_i)log(1-\pi(x_i))$$

再将其改为最小化负的对对数似然函数：

$$w^*=\underset{w}{argmin}\sum_{i=1}^{m}-y_ilog(\pi(x_i))-(1-y_i)log(1-\pi(x_i))$$

如此，就得到了Logistic回归的损失函数，即机器学习中的**「二元交叉熵」**（Binary crossentropy）：

$$J(w)=-\frac{1}{m}\sum_{i=1}^{m}y_ilog(\pi(x_i))+(1-y_i)log(1-\pi(x_i))$$

此时转变为以负对数似然函数为目标函数的最优化问题，采用梯度下降法进行优化。

## KL散度

KL散度这个概念知道的人可能相对极大似然估计更少一点，具体可以看机器学习笔记---信息熵。简单来说，**「KL散度是衡量两个概率分布的差异」**。

逻辑回归模型最后的计算结果（通过sigmoid或softmax函数）是各个分类的概率（可以看做是各个分类的概率分布）。那么假设真实的概率分布是$p(x)$，估计得到的概率分布是$q(x)$， 这两个概率分布的距离如何去衡量?

在信息论中，**「相对熵」**，也就是KL散度可以衡量两个概率分布的差异性。具体公式为:

$$D_{KL}(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} = \sum_x p(x)(logp(x) - logq(x))$$

并且简单转化，可以得到:

$$D_{KL}(p||q) = -\sum_x p(x)logq(x) - (-\sum_x p(x)logp(x)) = H(p,q) - H(p)$$

其中对于$H(p,q) = -\sum_x p(x)logp(x)$就是**「交叉熵」**，$H(p)$是真实分布的信息熵，所以

KL散度 = 交叉熵 - 真实概率分布的熵

因为交叉熵越大，KL散度越大，也可以用交叉熵来衡量两个概率分布之间的距离，所以逻辑回归使用交叉熵作为逻辑回归的损失函数。

## 总结

以上便是个人对这个问题的理解，如果解释有误的话，欢迎加我好友进行讨论。由于个人考虑到在一篇文章中包含过多的内容和公式可能会产生疲倦，所以只是集中于单个问题，对于其他逻辑回归相关的面试内容，也可以一起交流。

往期回顾

机器学习笔记---给“过拟合”下一个准确且规范的定义
机器学习笔记---正则化为什么可以抑制过拟合？
机器学习笔记---从极大似然估计的角度看待Logistic回归