开课吧数据竞赛第二课-钟老师-20191027

笔记本： 开课吧-小钟讲课
创建时间： 2019/10/21 10:44 **更新时间：** 2019/10/27 15:47
作者： 你看起来好像很好吃n_n
URL： https://blog.csdn.net/mathlxj/article/details/81490288

开课吧-数据竞赛及相关问题从小工到专家

时间：2019-10-27

1. 常用免费机器学习竞赛资源介绍

ti-one
ai_studio

```
import os
os.chdir('/cos_person/lxj/')
path = '../kaikeba/'
```

2. 二分类问题

概念：输入变量X和输出变量Y有不同的类型，可以是连续的，也可以是离散的。人们根据输入、输出变量的不同类型，对预测任务给与不同的名称：输入变量和输出变量均为连续变量的预测问题称为回归问题；输出变量为有限个离散变量的预测问题称为分类问题；输入变量与输出变量均为变量序列的预测问题称为标注问题。
二分类：

二分类问题就是简单的“是否”、“有无”问题，如下图我们判断是否为皮卡丘。我们如果提前预知下图为皮卡丘，那么通过我们的视神经系统能够很快分辨出下图是否为皮卡丘，但是对于机器来说分辨这张图却不是那么容易，更具体的来说机器只能读取这幅图的数字特征（如图像的大小，通道数等），在此我们以每个像素点的三原色对应的数值作为这幅图的数组特征。

二分类评价指标

1. 准确率
2. 混淆矩阵
3. 精准率,召回率，F1_score
4. auc
5. logloss

二分类算法:

1. Logistic回归
2. SVM
3. 决策树
4. 随机森林
5. Adaboost
6. xgboost
7. lightgbm
8. catboost
   9.朴素贝叶斯

1.1 线性回归原理

概念:
线性回归（Linear Regression）是一种通过属性的线性组合来进行预测的线性模型，其目的是找到一条直线或者一个平面或者更高维的超平面，使得预测值与真实值之间的误差最小化。
线性回归:

$$h(x) = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b$$

- 当只有一条x1时，只有h(x)为直线

- 当有两个值x1,x2两个变量的时候，h(x)为一个平面
- 当有更多变量时，h(x）为高维的。

线性回归是通过数据在N维空间找到h(x)来描述这些数据的规律，这是一个叫做拟合的过程，h(x)叫做拟合线。
h(x)的预测值会和真实值会有所偏差，真实统计和h(x)预测数据的差称为残差。残差有正的有负的，为了降低计算复杂性，我们使用这个差值的平方进行计算。为了获得最好的h(x),保证个点与实际数据的残差平方的总和最小。
代价函数为：

$$J = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$$

为了获得使得J最小的W和b：
主要有：

1. 偏导法
2. 正规方程法
3. 梯度下降 等等
   优缺点：
4. 权重W是每个*x*的权重，通过w的大小可以看出每个x的权重的大小，可以判断因子的重要性
5. 有很好的解释性
6. **缺点**：非线性数据拟合不好

1.2 逻辑回归原理

从1.1我们知道，从上面h(x)公式我们可以发现，h(x)预测值是连续的，所以这是一个回归模型。那如果我们希望输出的值是离散的，也就是预测值是离散值，所以线性回归是回归模型。那需要将h(x)进行一次函数转换，通过变成（
g(Y)。其中g(Y)某些值属于类别1，另一些g(Y)值属于类别，这样的模型则为二分类模型。二元逻辑回归就是这样来的。
此时函数g一般为为

$$g(z) = \frac{1}{1 + e^{-z}}$$

其中z=h(x)

有了 Sigmoid 函数之后，由于其值取值范围在[0,1]。

一般 Sigmoid 函数计算得到的值大于等于0.5的归为类别1，
小于0.5的归为类别0：

$P(y=1|x,\theta)=h\theta(x)$
$P(y=0|x,\theta)=1-h\theta(x)$
损失函数：
如果按线性回归的思想是：

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2m}\sum_{i=1}^{m}(\frac{1}{1+e^{-\theta^T x^{(i)}-b}} - y^{(i)})^2$$

但是这个函数不是凸函数，不好优化。

这是二分类模型，那么肯定是预测为真实为1的样本预测概率越接近1，损失越小；和预测为真实为0的样本预测概率越接近0，损失越小。

假设预测真实样本为1的概率为

pi，则预测真实样本为0的概率为

1 - pi。则预测概率为：

$$p(y_i) = p_i^{y_i} * (1 - p_i)^{1-y_i}$$

得到这个函数得到最大似然函数：

$$\prod_{i}^{N} h(x_i)^{y_i} * (1 - h(x_i))^{1-y_i}$$

两边取对数：

$$L(w) = \sum_i (y_i * logh(x_i) + (1 - y_i) * log(1 - h(x_i)))$$

$$= \sum y_i(logh(x_i) - log(1 - h(x_i))) + log(1 - h(x_i))$$

$$= \sum y_i log\frac{h(x_i)}{1 - h(x_i)} + log(1 - h(x_i))$$

$$= \sum y_i(w^T x_i) + log(1 - \frac{1}{1 + e^{-wx_i}})$$

$$= \sum_i (y_i * (w^T x_i) - log(1 + e^{w^T x_i}))$$

这里仅用随机梯度下降优化损失函数：

损失函数：

$$L(w) = \sum_i (y_i * (w^T x_i) - log(1 + e^{w^T x_i}))$$

对损失函数两边求导：

$$\frac{dL}{dw} = yx - \frac{1}{1+e^{w^T x}} * e^{w^T x} * x$$

$$= x(y - h(x))$$

最终迭代权值优化：

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) x_i^j$$

逻辑回归的优缺点优点：
1.容易理解和实现，可以观测样本的概率分数
2.训练速度快
3.由于经过了sigmoid函数的映射，对数据中小噪声的鲁棒性较好
4.不受多重共线性的影响(可通过正则化进行消除)缺点：
1.容易欠拟合
2.特征空间很大时效果不好
3.由于sigmoid函数的特性，接近0/1的两侧概率变化较平缓，中间概率敏感，波动较大；导致很多区间特征变量的变化对目标概率的影响没有区分度，无法确定临界值。

3. 二分类比赛快速实现

交叉验证(Cross Validation)为CV.CV是用来验证分类器的性能一种统计分析方法,基本思想是把在某种意义下将原始数据(dataset)进行分组,一部分做为训练集(train set),另一部分做为验证集(validation set),首先用训练集对分类器进行训练,在利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标.

Kfold
原始数据分成K组(一般是均分),将每个子集数据分别做一次验证集,其余的K-1组子集数据作为训练集,这样会得到K个模型,用这K个模型最终的验证集的分类准确率的平均数作为此K-CV下分类器的性能指标.

StratifiedKFold 是 k-fold 的变种，会返回 stratified（分层） 的折叠：每个小集合中， 各个类别的样例比例大致和完整数据集中相同。

1. 「二分类算法」提供银行精准营销解决方案
2. Titanic: Machine Learning from Disaster
3. 科大讯飞反欺诈