作业4 Logistic回归、SVM

1. 给定如下4个输入特征的4个样本，采用Logistic回归，

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **T1** | 5 | 3 | 1 | 1 | 1 |
| **T2** | 4 | 2 | 1 | 1 | 1 |
| **T3** | 2 | 1 | 2 | 3 | 0 |
| **T4** | 1 | 2 | 3 | 2 | 0 |

1. 初始化权重，，采用梯度下降，计算每个样本上的梯度（采用L2正则，）；

答：Logistic回归的目标函数为：

其中，

，

，

梯度为

**代入数值:**

，

，

类似的，

：

第一个样本的梯度：（b）

其余略

1. 假设当前得到参数估计为，。给定测试样本T5: ，给出样本的预测结果。

答：，所以预测其标签为0.

代码为：

import numpy as np

def sigmoid(x):

if x>0:

return 1.0/(1.0+np.exp(-x))

else:

return np.exp(x)/(1.0+np.exp(x))

x = [1, 3, 4, 2]

weights = [0.482, 0.179, -0.512, -0.524]

bias = 0.187

lr = np.dot(x, weights) + bias

y\_pred = sigmoid(lr)

1. 根据下表，绘制ROC曲线（阈值分别取0、0.2、0.4、0.6、0.8和1），并解释你是选择使用分类器1还是分类器2。

|  |  |  |
| --- | --- | --- |
| Class | 分类器1： | 分类器2： |
| P | 0.83 | 0.92 |
| N | 0.78 | 0.62 |
| P | 0.62 | 0.52 |
| N | 0.48 | 0.49 |
| N | 0.32 | 0.38 |
| N | 0.22 | 0.28 |

答：正样本数，负样本数

图表, 折线图

描述已自动生成

选择分类器1，因为分类器1具有更高的ROC曲线下面积（AUC）。这表明平均而言，对于相同的FPR值，它可以实现更高的TPR。

补充：ROC的绘制：

ROC的每个点对应一个阈值,而这个阈值对应了按照预测概率降序排序后样本的一种分类方式(排名第几之前的样例全部预测为正)。

排序后的样本从前往后遍历,依次将值作为阈值threshold，当测试样本属于正样本的概率大于或等于这个threshold时，我们认为它为正样本，否则为负样本。

举例来说，对于图中的第2个样本，其值为0.78，那么样本1，2，被认为是正样本，其他样本则都认为是负样本。每次选取一个不同的threshold，(，), (，)我们就可以得到一组FPR和TPR，即ROC曲线上的一点。

当我们将threshold设置为1和0时，分别可以得到ROC曲线上的(0.0)和(1,1)两个点。将这些FPR、TPR对连接起来，就得到了ROC曲线。当threshold取值越多，ROC曲线越平滑。

AUC的一般判断标准

0.5 - 0.7:效果较低，但用于预测股票已经很不错了

0.7-0.85:效果一般

0.85-0.95:效果很好

0.95-1:效果非常好，但一般不太可能

AUC的物理意义就是：**随机选出一对正负样本，模型对正样本的打分大于对负样本打分的概率**。

假设我们的测试集将正负样本按照模型预测得分从小到大排序，正样本总数为，负样本的总数为。对于第个正样本，假设它的排序为，那么说明排在这个正样本前面的总样本有 个，其中正样本有 个（因为这个正样本在所有的正样本里面排第 个），所以排在第个正样本前面（得分比它小）的负样本个数为个。也就是说，**对于第个正样本来说，其得分比随机取一个负样本大的概率是**，其中是负样本的总数。所以，平均下来，随机取得的正样本得分比负样本大的概率为：

其中是正样本的总数

假设正样本总数为，是负样本的总数，我们记排序后样本中第个负样例出现的位置为,那它之前的正样例数为。

这样ROC上的点从原点开始,分别是

(0,0),(FPR1,TPR1),(FPR2,TPR2),...,(FPRN,TPRN),(1,1)

等于

(0,0),,,...,,(1,1)

- 理想的二分类模型是排序后的样本中所有正样例排在所有负样例之前，ROC曲线就是连接 (0,0),(1,1) 的曲线。换句话说,对 ,.AUC就介于1和 .当负样例数无限大时,AUC就是1.

- 最糟糕的二分类器就是排序后的样本中所有负样例排在所有正样例之前,ROC曲线就是连接 (0,0),, (\frac{N-1}{N}, 0) 和(1,1)的折线.换句话说, 对 , AUC就介于 和0和 .当N负样例数无限大时,AUC就是0.

- 排序后的样本中第1个负样例出现的位置越靠前,召回率越低,这个点 (FPR1,TPR1)离x轴越近,ROC曲线下的面积越小。同样,第2个负样例出现的位置越靠前,召回率越低,ROC曲线第二个点下面的面积越小.推而广之,全部负样例在排序后的样本中出现的位置越靠前,ROC曲线覆盖的面积越小,也就是AUC越小。全部样例排序后，样本位置和AUC的具体关系为：

1. 给定下列3个样本点，请计算硬间隔SVM分类器。

|  |  |
| --- | --- |
|  |  |
| 3 | 1 |
| 1 | 1 |
|  |  |

方法一：画图，可以看出，第2个点和第3个点是支持向量。 ,

**方法二：求解原问题**

变成标准形式：

基于该标准形式，构造的拉格朗日函数为

上面等式给出了2个方程，还需要加上松弛条件：，。

第2个点和第3个点是支持向量，，：

，，最大间隔超平面为，。

但还需要验证，满足约束条件，所以该解是最优解。

**方法三：求解对偶问题**

先求解，、都与相关，得到了 也就求得了超平面参数。

|  |  |
| --- | --- |
|  |  |
| 3 | 1 |
| 1 | 1 |
|  |  |

，，，

，，，

由，得到将，将其带入目标函数

由，得到，代入。易知时取极大值。但是注意到这里的 不满足约束条件 ，所以此时极大值就不在边界内而是在边界上，需分别考虑 为 0 的情况：

1. ，，，；
2. ，，。

综上两种情况，的极大值为，对应的，，。

通过拉格朗日乘子的特点就可以知道 、为支持向量（因为，） ，所以约束条件  ，，所以 、为是在间隔边界上的。

然后再代入 、的计算公式，求得：

结果同原问题求解相同。

1. 软间隔SVM分类器通过引入松弛变量来放松优化约束，允许在分类中出现错误。原始形式的软间隔SVM分类器如下：

以下说法是否正确？并给出理由。

1. 增加超参数倾向于减少训练误差。

正确，高给出错误分类实例惩罚更高，导致错误分类更少。

1. 增加超参数往往会降低间隔

正确。高导致更少的错误分类，并且通常需要更小的间隔来减少错误分类

1. 硬间隔SVM是超参数设置为0的软间隔的特殊情况。

错误，它是具有无限的软间隔，因此不允许错误分类

（4） 增加超参数往往会降低对异常值的敏感性。

错误，增加会增加对异常值的敏感性，因为错误分类的实例会受到更高的惩罚。

1. 考虑以下4个由不同核函数和/或松弛惩罚的导致的SVM决策边界（a）-（d）。标签的两类训练数据分别用三角形和正方形表示。支持向量被描述为实心三角形和正方形。请将每个决策边界与下面（A）-（D）中最可能的优化设置相匹配，并简要说明配对的理由。

图示

描述已自动生成

1. 的软间隔线性SVM。

（c） 线性决策边界，较小，允许在间隔内进行一些错误分类

1. 的软间隔线性SVM。

（a）线性决策边界，较大，不太允许错误分类，在间隔内没有支持向量

1. 核函数为的硬间隔SVM。

（b） 具有更宽核（的RBF核）的非线性决策边界

（D） 核函数为的硬间隔SVM。

（d） 具有较窄核（的RBF核）的非线性决策边界