作业4 Logistic回归、SVM、统计学习理论

作业提交截止日期：11月29日 23:55

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；

答：Logistic回归的目标函数为：（采用L2正则，）；

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. 假设次梯度下降迭代后，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。

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，将其代入目标函数

因此，得到。

又由于，不能是不同符号，所以极大值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。

答：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

答：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。

答：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。

答：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. 假设我们采用不同的验证集划分方式： 2折交叉验证、10折交叉验证、留一交叉验证、单次70%/30%的训练集/验证集划分。
2. 不同的验证集划分方式会对模型性能产生什么影响？（如训练误差、泛化误差）

答：从2折交叉验证、10折交叉验证、留一交叉验证，训练样本数目越来越多，得到的训练误差和泛化误差更接近采用所有数据训练模型时的性能。

1. 哪种方式得到的验证误差会提供 在“未见过的测试集”上误差的最佳近似？

答：一般而言，留一交叉验证的验证误差是测试误差的好估计，但留一交叉验证需要重复次的模型训练，每次训练的训练数据多，运行成本高。（为数据集中的样本数目）

1. 原始数据集有多大有影响吗？对于一个非常大或非常小的数据集，你会得到不同的结论吗？

答：原始数据集大小有影响。对非常大的数据集，上述方法会得到的相似的结果。但一般选择单次留出验证集的方式，只需训练一次模型，运行成本低，此时训练数据充分，验证集也足够多，验证误差为测试误差的很好估计；对非常小的数据集，留一交叉验证才能得到足够多的训练数据，此时验证误差是测试误差的较好估计。

1. 就计算而言，哪种方式最快？

答：单次的70%/30%的训练集/验证集划分最快。