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；
2. 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。
3. 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 |

1. 给定下列3个样本点，请计算硬间隔SVM分类器。

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

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。

（2） 增加超参数往往会降低间隔

（3） 硬间隔SVM是超参数设置为0的软间隔的特殊情况。

（4） 增加超参数往往会降低对异常值的敏感性。

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）
3. 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？
4. 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？
5. 就计算而言，哪种方式最快？