CSDN
不 止 于 代 码

2.6 Logistic回归之模型选择 / 参数调优

CSDN学院

2017年11月

大纲

- Logistic回归基本原理
- 多类Logistic回归
- Scikit learn 中的Logistic回归实现
- 分类模型的评价
- 模型选择与参数调优
- 案例分析

网格搜索（Grid Search）

- 不同超参数下的模型性能不同。
- 为了找到最佳模型，通常对这些超参数设定搜索范围
- 多个超参数可以联合一起优化，得到超参数的搜索网格
  - 如：LogisticRegression中的超参数*penalty*和*C*一起优化
    - *penalty*可取 ‘l2’ 或 ‘l1’
    - *C*假设取值范围为： 0.001, 0.01, 0.1, 0, 1, 10, 100, 1000
  - 则搜索网格为：

‘l2’	0.001	0.01	0.1	1	10	100	1000
‘l1’	0.001	0.01	0.1	1	10	100	1000

LogisticRegression超参数调优

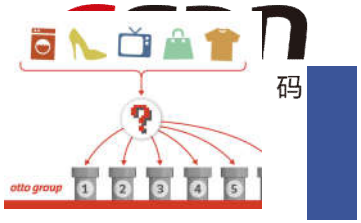
- 超参数调优需先确定超参数的搜索网格，然后对每个可能的超参数组合评估其性能。
- 对LogisticRegression的超参数调优，scikit learn提供给两种实现方式：
  - 1. 同其他estimator一样，调用GridSearchCV（集成了网格搜索和交叉验证）：设置候选参数集合、根据候选参数集合构造GridSearchCV、调用GridSearchCV 的fit函数；
  - 2. LogisticRegressionCV 类内置的LR的交叉验证，用于找到最优的*C*参数

LogisticRegressionCV

- LogisticRegressionCV 使用了内置的Logistic回归的交叉验证，用于找到最优的*C*参数。（正则参数 penalty可设为 ‘l1’ 或 ‘l2’ ）

- 对于多分类问题
  - 如果multi_class参数设置为 “ovr”，对于每个类都获得一个最优的*C*；
  - 如果multi_class设置为” multinomial”, 将获得一个最优的C，它使得交叉熵的loss（ corss-entropy loss ）最小。

案例分析：
Otto Group Product Classification Challenge



- 竞赛官网：https://www.kaggle.com/c/otto-group-product-classification-challenge
- 电商商品分类：
  - Target：共9个商品类别
  - 93个特征：整数型特征

CSDN
不 止 于 代 码

THANK YOU

AI 100