CSDN
不 止 于 代 码

1.6 线性回归模型
——模型选择

CSDN学院

2017年10月

线性回归

- 模型
  - 目标函数（损失函数、正则）
    - 概率解释
- 优化求解
- 模型选择

线性回归模型

- 无正则的最小二乘线性回归（Ordinary Least Square，OLS）

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i \right)^2$$

- L2正则的岭回归（Ridge Regression）模型：

$$J(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i \right)^2 + \lambda \|\mathbf{w}\|_2^2$$

- L1正则的Lasso模型：

$$J(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i \right)^2 + \lambda |\mathbf{w}|$$

模型评估与模型选择

- 模型训练好后，需要在校验集上采用一些度量准则检查模型预测的效果
  - 校验集划分（train_test_split、交叉验证）
  - 评价指标（sklearn.metrics）
- 模型选择：选择预测性能最好的模型
  - 模型中通常有一些超参数，需要通过模型选择来确定
    - 线性回归模型中的正则参数λ
    - OLS中的特征的数目
  - 参数搜索范围：网格搜索（GridSearch）
- Scikit learn将交叉验证与网格搜索合并为一个函数：sklearn.model_selection.GridSearchCV

评价准则

- 模型训练好后，可用一些度量准则检查模型拟合的效果
  - 开方均方误差（rooted mean squared error，RMSE）：$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$
  - 平均绝对误差（mean absolute error，MAE）：$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$
  - R2 score：既考虑了预测值与真值之间的差异，也考虑了问题本身真值之间的差异（ scikit learn 线性回归模型的缺省评价准则）

$$SS_{res} = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2, SS_{tot} = \sum_{i=1}^{N}(y_i - \bar{y})^2, R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- 也可以检查残差的分布
- 还可以打印预测值与真值的散点图

Scikit learn中的回归评价指标

Regression	
‘explained_variance’	metrics.explained_variance_score
‘neg_mean_absolute_error’	metrics.mean_absolute_error
‘neg_mean_squared_error’	metrics.mean_squared_error
‘neg_mean_squared_log_error’	metrics.mean_squared_log_error
‘neg_median_absolute_error’	metrics.median_absolute_error
‘r2’	metrics.r2_score

线性回归中的模型选择

sklearn.model_selection

- Scikit learn中的model selection模块提供模型选择功能
  - 对于线性模型，留一交叉验证（*N*折交叉验证，亦称为leave-one-out cross-validation，LOOCV）有更简便的计算方式，因此Scikit learn提供了RidgeCV类和LassoCV类实现了这种方式
  - 后续课程将讲述一般模型的交叉验证和参数调优GridSearchCV

RidgeCV

- RidgeCV中超参数λ用alpha表示
- RidgeCV(*alphas=(0.1, 1.0, 10.0), fit_intercept=True, normalize=False, scoring=None, cv=None, gcv_mode=None, store_cv_values=False)*

```
from sklearn.linear_model import  RidgeCV

alphas = [0.01, 0.1, 1, 10,20, 30, 50, 60, 80,100]

reg = RidgeCV(alphas=alphas, store_cv_values=True)
reg.fit(X_train, y_train)
```

LassoCV

- LassoCV的使用与RidgeCV类似

- Scikit learn还提供一个与Lasso类似的LARS（least angle regression，最小角回归），二者仅仅是优化方法不同，目标函数相同。
- 当数据集中特征维数很多且存在共线性时，LassoCV更合适。

► 小结：线性回归之模型选择

- 采用交叉验证评估模型预测性能，从而选择最佳模型
  - 回归性能的评价指标
  - 线性模型的交叉验证通常直接采用广义线性模型的留一交叉验证进行快速模型评估
    - Scikit learn中对RidgeCV和LassoCV实现该功能