CSDN
不 止 于 代 码

3.2 随机森林

CSDN学院

2017年11月

► 随机森林（Random Forest）

CSDN
不 止 于 代 码

- 回归树算法的缺点之一是高方差
- 一种降低算法方差的方式是平均多个模型的预测：**Bagging**
  - **B**ootstrap **agg**regat**ing**
- 随机森林：Bagging多棵树

Bootstrap

- 通过从原始的N个样本数据$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$进行$N$次有放回采样$N$个数据$\mathcal{D}$’，得到bootstrap样本
  - 对原始数据进行有放回的随机采样，抽取的样本数目同原始样本数目一样

- 如：若原始样本为$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
- 则bootstrap样本可能为
  - $\mathcal{D}^1 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_4, \mathbf{x}_5\}$
  - $\mathcal{D}^2 = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5\}$

Bagging

CSDN
不 止 于 代 码

- 对给定有N个样本的数据集 $\mathcal{D}$进行Bootstrap采样，得到$\mathcal{D}^1$，在$\mathcal{D}^1$上训练模型$\hat{f}^1$
- 上述过程重复B次，得到B个模型，则B个模型的平均为

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

aggregating

- 可以证明（略）：Bagging可以降低模型的方差

随机森林（Random Forest）

CSDN
不 止 于 代 码

- 由于只是训练数据有一些不同，对回归树算法进行Bagging得到的多棵树高度相关，因此带来的方差减少有限
- 随机森林通过
  - 随机选择一部分特征
  - 随机选择一部分样本
- 降低树的相关性

- 随机森林在很多应用案例上被证明有效，但牺牲了可解释性
  - 森林：多棵树
  - 随机：对样本和特征进行随机抽取

Scikit learn中的Random Forest实现

CSDN
不 止 于 代 码

- sklearn.ensemble.RandomForestClassifier(*n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False, class_weight=None*)

► 小结

CSDN
不 止 于 代 码

- 随机森林：bagging多棵树
  - 树的数目
  - 树的复杂度

CSDN
不 止 于 代 码

THANK YOU

AI100