

3. 特征工程

1. Filter: 过滤法

1) Filter: 过滤法

根据每个属性的一些指标 (如方差), 来评估这个属性的重要程度, 然后对所有属性按照重要程度排序, 从高到低的选择属性。

2. Embedding: 嵌入式

2) Embedding: 嵌入式

把特征选择的过程作为学习过程的一部分, 在学习的过程中进行特征选择。即先使用某些机器学习算法和模型进行训练, 得到各个特征的权重系数, 根据权重系数从大到小选择特征。这些权重系数往往代表了特征对于模型的贡献或重要性。(决策树)

3. Wrapper: 包裹式

3) 包裹式 Wrapper

也是特征选择与算法训练同时进行的办法, 可以调用 `coef` 或 `feature-importance` 属性来完成特征选择。

但不同的是, 我们往往使用一个评估器作为黑盒来帮助我们选取特征。包裹法在初始特征集上训练评估器, 并且通过 `coef` 属性或 `feature-importance` 属性获得每个特征的重要性。然后, 从当前的一组特征中修剪最不重要的特征。在修剪的集合上递归地重复该过程, 直到最终到达所需数量的特征。

区别于过滤法和嵌入式的一次训练解决所有问题, 包裹法要使用特征子集进行多次训练, 因此包裹法成本较高。

最典型的方法：递归特征消除法 (RFE).

它是一种贪心的优化算法，旨在找到性能最佳的特征子集。它反复创建模型，并在每次迭代时保留最佳特征或剔除最差特征，下一次迭代时，它会使用上次建模中没有被选中的特征来构建下一个模型，直到所有特征都被剔除。然后，根据保留或剔除特征的顺序来对特征进行排名，



扫描全能王 创

Date: / / Page:

最终选出一个最佳子集。

包裹法是所有特征选择方法中最有利于提升模型表现的，它可以使用很少的特征达到很优秀的效果。但其计算量大，不太适合太大型的数据。