## 0 特征工程的定义

特征工程是机器学习应用的一个重要步骤，它在建模之前进行，是根据原始数据的特征，构造出任务和模型所需要的最能反映样本的特性的特征。比如对于分类任务，如果我们使用线性分类器，那么就希望样本的特征能尽量线性可分。比如我们利用线性分类器，基于关键词作情感分析，那么具有线性性质的词向量就是比较好的特征。

在很多机器学习任务中，特征工程对于应用效果有着决定性的影响。如果能够构造出特别合适的特征，那么模型的选择其实就不那么重要了。

## 一 特征工程的方法

特征工程是一个具体问题具体分析的玩意。对于不同的任务，特征工程的做法可能有很大的差别。很多时候，构造出好的特征需要很多的先验知识，需要对业务有比较深刻的理解。所以有的人说，“特征工程是一门艺术”。

这里很难穷举实际中各种任务的特征工程的最佳做法，仅介绍一个最基本的思路：特征选择。

特征选择，顾名思义，就是从原始数据中选择一些对任务来说最有价值的特征。其目的是去除无关特征和冗余特征。无关特征就是对于预测来说不起作用的特征；冗余特征就是与其他特征相关性非常强的特征。

与数据降维一样，特征选择也是压缩数据的特征维数。但不同的是，数据降维的思路是构造出不同于原始特征的一组新的特征，而特征选择是从原始的特征中选择一些特征。

特征选择的常用做法：

(1) 贪心选择：需要与模型绑定。

(2) 基于相关性计算：计算每个特征跟预测结果的相关性，相关性越强的特征越好。

(3) 基于L1正则：L1的稀疏作用相当于特征选择。

(4) 决策树算法：越靠近根节点的特征越好。