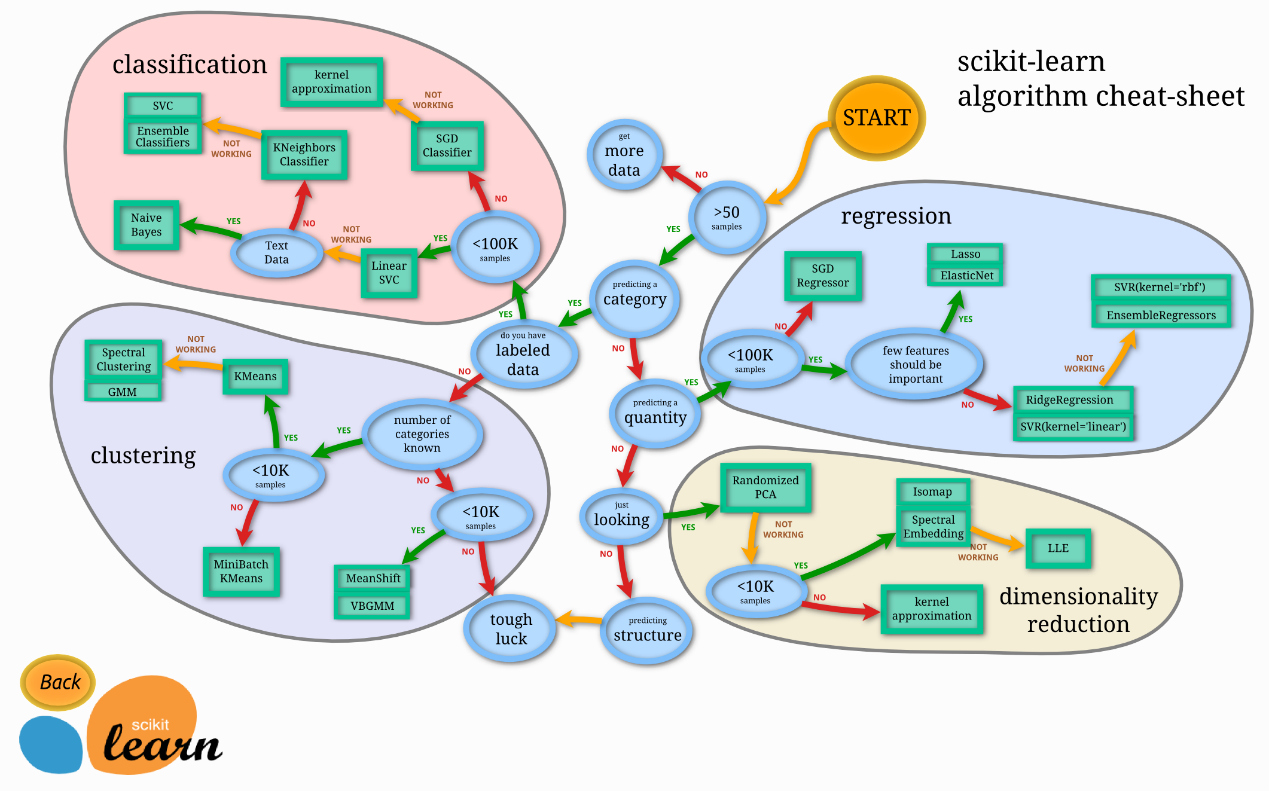
Sklearn



使用机器学习算法和模型进行数据挖掘，有时难免事与愿违：我们依仗对业务的理解，对数据的分析，以及工作经验提出了一些特征，但是在模型训练完成后，某些特征可能“身微言轻”——我们认为相关性高的特征并不重要，这时我们便要反思这样的特征提出是否合理；某些特征甚至“南辕北辙”——我们认为正相关的特征结果变成了负相关，造成这种情况很有可能是抽样与整体不相符，模型过于复杂，导致了过拟合。然而，我们怎么判断先前的假设和最后的结果之间的差异呢？

　　线性模型通常有含有属性coef\_，当系数值大于0时为正相关，当系数值小于0时为负相关；另外一些模型含有属性feature\_importances\_，顾名思义，表示特征的重要性。根据以上两个属性，便可以与先前假设中的特征的相关性（或重要性）进行对比了。但是，理想是丰满的，现实是骨感的。经过复杂的特征转换之后，特征矩阵X已不再是原来的样子：哑变量使特征变多了，特征选择使特征变少了，降维使特征映射到另一个维度中。