一个完整机器学习项目流程总结

展开

1. 实际问题抽象成数学问题

这里的抽象成数学问题，指的我们明确我们可以获得什么样的数据，

目标是一个分类还是回归或者是聚类的问题，如果都不是的话，如果划归为其中的某类问题。

2. 获取数据

获取数据包括获取原始数据以及从原始数据中经过特征工程从原始数据中提取训练、测试数据。

机器学习比赛中原始数据都是直接提供的，但是实际问题需要自己获得原始数据。“ 数据决定机器学习结果的上限，而算法只是尽可能的逼近这个上限”，可见数据在机器学习中的作用。总的来说数据要有具有“代表性”，对于分类问题，数据偏斜不能过于严重，不同类别的数据数量不要有数个数量级的差距。不仅如此还要对评估数据的量级，样本数量、特征数量，估算训练模型对内存的消耗。如果数据量太大可以考虑减少训练样本、降维或者使用分布式机器学习系统。

3. 特征工程

特征工程包括从原始数据中特征构建、特征提取、特征选择。特征工程做的好能发挥原始数据的最大效力，往往能够使得算法的效果和性能得到显著的提升，有时能使简单的模型的效果比复杂的模型效果好。数据挖掘的大部分时间就花在特征工程上面，是机器学习非常基础而又必备的步骤。数据预处理、数据清洗、筛选显著特征、摒弃非显著特征等等都非常重要。

4. 训练模型、诊断、调优

模型诊断中至关重要的是判断过拟合、欠拟合，常见的方法是绘制学习曲线，交叉验证。

通过增加训练的数据量、降低模型复杂度来降低过拟合的风险，提高特征的数量和质量、增加模型复杂来防止欠拟合。

诊断后的模型需要进行进一步调优，调优后的新模型需要重新诊断，这是一个反复迭代不断逼近的过程，需要不断的尝试，

进而达到最优的状态。

5. 模型验证、误差分析

通过测试数据，验证模型的有效性，观察误差样本，分析误差产生的原因，往往能使得我们找到提升算法性能的突破点。误差分析主要是分析出误差来源与数据、特征、算法。

6. 模型融合

提升算法的准确度主要方法是模型的前端（特征工程、清洗、预处理、采样）和后端的模型融合。

在机器学习比赛中模型融合非常常见，基本都能使得效果有一定的提升。

7. 上线运行

这一部分内容主要跟工程实现的相关性比较大。工程上是结果导向，模型在线上运行的效果直接决定模型的成败。

不单纯包括其准确程度、误差等情况，还包括其运行的速度(时间复杂度)、资源消耗程度（空间复杂度）、稳定性是否可接受。

*值得注意的是，以上流程只是一个指导性的机器学习流程经验，并不是每个项目都包含完整的流程。*