机器学习是什么，什么流程，需要哪些基础：

|  |
| --- |
| **可解决什么问题**：  1: 分类问题: 根据数据样本上抽取出的特征，判定其属于有限个类别中的哪一个  (**邮件识别**：垃圾或者正常，新闻分类)  2: 回归问题: 根据数据样本上抽取出的特征，预测一个连续值的结果  (电影票房，房价估计)  3:聚类问题: 根据数据样本上抽取出的特征，让样本抱抱团(相近/相关的样本在一团内)  (google 新闻分类)  基础：数学 ＋ 机器学习典型方法 ＋ 编程  **微积分**：理解微分的集合意义，优化理论，例如剃度下降，牛顿发，怎样调参数，怎样得到快的收敛速度，凸优化理论，KKT条件，对偶性等等，  **线性代数**：四个空间， 投影矩阵啊，最小二乘拟合的理解，PCA和SVD  概率论：ML,MAP,贝叶嘶，隐形马尔可夫模型，高斯混合模型  **典型的算法**：   1. 处理分类问题的常用算法包括：逻辑回归(工业界最常用)，支持向量机，随机森林，朴素贝叶斯(NLP中常用)，深度神经网络(视频、图片、语音等多媒体数据中使用)。 2. 处理回归问题的常用算法包括：线性回归，普通最小二乘回归（Ordinary Least Squares Regression），逐步回归（Stepwise Regression），多元自适应回归样条（Multivariate Adaptive Regression Splines） 3. 处理聚类问题的常用算法包括：K均值（K-means），基于密度聚类，LDA等等。 4. 降维的常用算法包括：主成分分析（PCA）,奇异值分解（SVD） 等。 5. 推荐系统的常用算法：协同过滤算法 6. 模型融合(model ensemble)和提升(boosting)的算法包括：bagging，adaboost，GBDT，GBRT 7. 其他很重要的算法包括：EM算法等等。   **机器学习项目的基本流程**：  1:**抽象成数学问题**：  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**：对于数据，要注意以下几个点  1数据要有代表性，否则必然会过拟合  2考虑数据大小和内存的问题，估算出其对内存的消耗程度，判断训练过程中内存是否能够放得下。如果放不下就得考虑改进算法或者使用一些降维的技巧了。如果数据量实在太大，那就要考虑分布式了  3:**特征预处理和特征选择**：  筛选出显著特征、摒弃非显著特征，需要机器学习工程师反复理解业务。这对很多结果有决定性的影响。  特征选择好了，非常简单的算法也能得出良好、稳定的结果。这需要运用特征有效性分析的相关技术，如相关系数、卡方检验、平均互信息、条件熵、后验概率、逻辑回归权重等方法。  4:**训练模型和调优**  直到这一步才用到我们上面说的算法进行训练  5: **模型诊断**：确定模型调优的方向与思路  1:过拟合、欠拟合 判断是模型诊断中至关重要的一步。常见的方法如交叉验证，绘制学习曲线等。过拟合的基本调优思路是增加数据量，降低模型复杂度。欠拟合的基本调优思路是提高特征数量和质量，增加模型复杂度。  2: 误差分析 也是机器学习至关重要的步骤。通过观察误差样本，全面分析误差产生误差的原因:是参数的问题还是算法选择的问题，是特征的问题还是数据本身的问题  注意：诊断后的模型需要进行调优，调优后的新模型需要重新进行诊断，这是一个反复迭代不断逼近的过程，需要不断地尝试， 进而达到最优状态  6: **模型融合**：一般来说，模型融合后都能使得效果有一定提升。而且效果很好 |