**机器学习面试题目总结**

**一.** **特征选择和降维**

**特征选择**：原有特征选择出子集，不改变原来的特征空间

**降维**：将原有的特征重组成为包含信息更多的特征，改变了原有的特征空间

**降维的主要方法**

Principal Component Analysis(主成分分析)

Singular Value Decomposition(奇异值分解)

**特征选择的方法**

1. Filter方法：根据单个特征对目标的重要性选择最佳特征

Chi-squared test(卡方检验)

correlation coefficient scores(相关系数)

互信息/最大信息系数

1. Wrapper方法：

基本思想：对特征组合进行评价

通常以递归特征删除的方式实现

* + 用所有特征训练模型，得到特征系数（如线性模型）或特征重要性（如xgboost等基于树的模型）
  + 根据特征系数或特征重要性，每次删除最不重要的（5-10%）的特征，观察模型性能变化，直至模型性能出现大的下滑或达到特征数目

1. Embedded方法

基本思想：使用所有特征训练模型，但模型训练结束后，模型本身已经选择好了特征

**L1正则：**L1正则方法具有稀疏解的特性，因此天然具备特征选择的特性。但是要注意，L1没有选到的特征不代表不重要，原因是两个具有高相关性的特征可能只保留了一个。

GBDT（其他基于树的模型）: 树模型在每次分裂选择一个特征根据阈值进行分裂，模型没有选择其进行分裂的特征在模型中不起作用。

**二. 机器学习中，有哪些特征选择的工程方法？**

**数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已**

1. 问题一中的特征选择方法
2. 去掉方差太小的特征
3. 通过特征组合后，再选择特征

可分步执行，先从单个特征层面，选择一些重要特征，然后对这些特征进行组合，对组合后的特征再进行特征选择

也可特征组合和特征选择一起进行（矩阵分解FM/FFM）

1. **通过深度学习来进行特征选择**

**三、机器学习中的正负样本均衡**

**在分类问题中，**。

**四、线性分类器与非线性分类器的区别及优劣**

**区别：**所谓线性分类器即用**一个超平面将正负样本分离开，表达式为** *y***=w**T**x 。这里是强调的是平面。**

而非线性的分类界面没有这个限制，**可以是曲面，多个超平面的组合等。**

**典型的线性分类器有：**LDA（线性判别分析，课程内容没讲），Logistic回归，SVM（线性核）；

**典型的非线性分类器有：**kNN（k近邻），决策树（和基于树的集成学习模型），SVM（非线性核）

**优缺点**：

1. 线性分类器判别简单（测试）、易实现、需要的计算量和存储量小。
2. 如果任务本身是非线性的，非线性分类器通常会比线性分类器表现得更精准。
3. 线性模型的非线性化：
   1. 特征🡪对原始特征进行非线性扩展，再用线性判别模型
   2. 对分段线性模型组合，逼近全局的非线性
   3. 线性模型核化（核技巧）

**五. 如何解决过拟合问题？**

**过拟合：**模型在训练集表现好，在测试数据表现不好，即模型的泛化能力不够。从另外一个方面来讲，模型在达到经验损失最小时，模型复杂度较高，结构风险没有达到最优。

**解决：**

1. 模型：限制模型的复杂性，对模型增加正则项，使得机器学习学得不那么彻底（如训练数据拟合得不要太好），从而降低机器学到局部特征和错误特征的概率
2. 数据：数据去燥、样本增多、特征选择
3. **L1和L2正则的区别，如何选择L1和L2正则？**

**L1优点**是能够获得sparse模型，对于large-scale的问题来说这一点很重要，因为可以减少存储空间。**缺点是**加入L1后目标函数在原点不可导，优化时需要做特别处理（如次梯度）。

**L2优点**是实现简单，能够起到**正则化**的作用**。缺点就是L1的优点：无法获得sparse模型**。

实际上要获得真正sparse的模型，要用L0正则化。L1正则也是对L0正则的近似。

参考：机器学习中的范数规则化之（一）L0、L1与L2范数

**七、有监督学习和无监督学习的区别**

**有监督学习**：对有标签的训练样本进行学习，以尽可能对训练样本集外的数据进行标签预测。

**无监督学习：**对没有概念标记的训练样本进行学习，以发现训练样本集中的结构性知识。聚类就是典型的无监督学习

练习题：

1. 每个实体有不同属性，现有很多实体的各种属性数据，如何判断两个实体是否是同一种东西？
2. L1损失和L2损失的区别
3. 推荐系统的算法中最近邻和矩阵分解各自适用场景
4. 多分类怎么处理
5. 为什么会产生过拟合，有哪些方法可以预防或克服过拟合？
6. 给你公司内部群组的聊天记录，怎样区分出主管和员工？
7. 路段平均车速反映了路况，在道路上布控采集车辆速度，如何对路况做出合理估计？采集数据中的异常值如何处理？
8. 如何根据语料计算两个词词义的相似度？
9. 在百度贴吧里发布 APP 广告，问推荐策略？
10. 如何判断自己实现的 LR、Kmeans 算法是否正确？