* **有哪些特征选择的方法？**

1. 计算每一个特征与响应变量的相关性：工程上常用的手段有计算**皮尔逊系数**和**互信息系数**，皮尔逊系数只能衡量线性相关性而互信息系数能够很好地度量各种相关性，但是计算相对复杂一些，好在很多toolkit里边都包含了这个工具（如sklearn的MINE），得到相关性之后就可以排序选择特征了；

互信息：设两个随机变量https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D41/sign=57d638070ff431adb8d242384a36d8b5/9345d688d43f879441a395bbd11b0ef41ad53ac3.jpg的联合分布为https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D46/sign=d3d83c2dd73f8794d7ff4928d31b7a75/622762d0f703918fe171e25f523d269758eec487.jpg，边际分布分别为https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D70/sign=bd7d3a6da486c9170c035039c83d6f09/8601a18b87d6277fb698ecad2b381f30e824fc99.jpg，互信息https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D51/sign=9d8320b64a540923ae69637f93580985/b219ebc4b74543a9921424591d178a82b9011413.jpg是联合分布https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D46/sign=bec41fbabd096b6385195f560d33e45f/0823dd54564e9258658b5e7b9f82d158cdbf4eec.jpg与乘积分布https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D64/sign=e5d39e6a38c79f3d8be1e734b8a169c6/1e30e924b899a90160e680a31e950a7b0308f550.jpg的相对熵，即

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D260/sign=81f9bba74f086e066ea8384d32097b5a/eaf81a4c510fd9f9ee4a72ff262dd42a2934a4c2.jpg

它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。

1. **构建单个特征的模型，通过模型的准确性为特征排序，借此来选择特征，**另外，记得JMLR'03上有一篇论文介绍了一种基于决策树的特征选择方法，本质上是等价的。当选择到了目标特征之后，再用来训练最终的模型；
2. **通过L1正则项来选择特征：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，L**1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，如果要确定哪个特征重要应再通过L2正则方法交叉检验**；
3. 训练能够对特征打分的预选模型：RandomForest和Logistic Regression等都能对模型的特征打分，通过打分获得相关性后再训练最终模型；

* **数据预处理的流程？**

为什么要进行数据预处理？真实世界中，数据往往比较脏，主要表现在数据不完整，噪声、特征表示不一致、特征冗余或重复等…

预处理的流程，一般包括：数据清洗、数据集成、数据转换、数据约减、数据离散化等过程。

**1、统计计算数据的描述性信息：**

统计数据的均值、方差、最值、分位数、噪声点、分布、散点图等等，分析数据的特征。

**2、数据清洗：**

* + **处理数据缺失：**（1）丢弃缺失的数据/丢弃缺失的属性，（2）使用均值、中值/插值方法进行填充,（3）不处理 ---- 某些模型可以接收缺失值
* **处理噪声/离群点：**
  + - **筛选**：（1）根据先验只是设置阈值筛选，（2）箱线图，（3）聚类。
    - **处理**：（1）分箱处理：等宽度、等深度，（2）使用回归曲线进行平滑，（3）聚类移除离群点，（4）人工筛选

**3、数据规范化**

将数据变换到特定的范围，为了消除特征数据之间的量纲影响，因为不同的数据在不同列数据的数量级相差过大的话，计算起来大数的变化会掩盖掉小数的变化。

另一方面，进行归一化可以加速收敛。

（1）最大最小值：也称为归一化，受噪声影响大，适合精确的小数据场景，（2）z-score：也称为标准化，适合噪声较大的大数据场景

**4、数据集成**

将不同的数据源的数据合并，需要处理数据在表示方法、尺度等方面的区别

**5、特征筛选/降维**：

**筛选**：去除冗余特征 — 计算相关系数、卡方值等，使用模型进行选择…

**降维**：PCA等选择主成分…

**6、连续数据离散化：**

连续数据离散化有很多好处

* **处理过拟合的手段？**

过拟合是由于模型过于复杂拟合了噪声的特性，导致模型在训练数据上效果很好，但是在测试数据上效果很差。

1. **增加训练样本数量**：

采集更多的样本

数据增强—人工生成样本

1. **选择合适的模型结构**

一方面可以选择合适的模型

另一方面，对于某个模型，可以设置合适的模型参数，保证模型复杂度与样本数量匹配

1. **提前结束训练**

可以只保存最佳的模型，或者设置一个准确率门槛，达到设定的值提前结束训练。

1. **使用正则化约束**

正则化约束使得模型的复杂度与准确率之间得到一个制衡，我们需要的是在二者之间取一个折衷，常用L1,L2

1. **非常规手段：**

不同的模型，可能有不同的防止过拟合的手段，例如：

决策树：剪枝

神经网络：dropout

* **连续特征离散化哑变量编码的好处？**

一般连续特征可以进行离散化之后，进行二进制编码。

* + 可以平滑一些噪声的影响，对异常数据具有更强的鲁棒性；
  + 离散化将一个特征变为多个，特征的维度增加了，可以引入非线性，让样本更容易区分，模型更简单；
  + 有些模型需要属性特征

**处理数据不均衡的手段？**

处理数据不均衡问题的处理，目前主要从三个方面进行考虑：

* 首先是从样本层面进行考虑，**设计合理的采样方法**，使得训练数据均衡；
* 其次是从模型和算法的角度考虑，**设计或改进方法，减小不均衡的影响**；
* 最后是从分类的评价标准上考虑，**设计合理的评价指标**，评估模型分类非均衡数据的效果。

**采样层面：**

* **随机欠采样和过采样**：前者减少多类，后者减少少类；
* **依据信息的欠采样**：从多类选择有代表性的样本；
* **基于数据生成的综合过采样：**SMOTE方法，在少类中，随机选择样本，加上其k近邻与系数的乘积。

**算法层面：**

* 选择**不容易受到数据不均衡影响的模型**：如决策树、随机森林、SVM等
* 设计**代价敏感的方法**，其做法就是设置权重，让错分少类的代价更高。

**评价指标层面：**

传统意义上，我们就简单用**分类准确率**来描述，但是其能展示的信息很有限。

对于不同的问题，我们可以使用不同的度量来反映模型的性能：

**精度**：Precision = TP / (TP + FP)，表示所有分到正类的样本中，正确分类的比例

**召回率/查全率**：Recall = TP / (TP + FN)， 表示所有的正样本，被正确分类的比例

**F1值**： 2/F1 = 1/Precision + 1/Recall，它是精度和召回率的调和平均

**ROC曲线**：Recall和1-Sensitiy的曲线，即正样本被分为正的比例与负样本被分为正的比例之间的曲线关系。曲线中的点的变化受到模型阈值设定的影响。（如Logistic回归，认为概率大于0.3为正）ROC曲线偏离对角线越远，模型效果越好

**AUC值**：ROC曲线与x轴围成的面积，即ROC曲线下的面积，面积越大，模型的性能越好