# 任务与评估

阶段：数据样本、特征工程、特征表示、度量、任务、训练、预测

回归 分类 多酚类 评价函数 auc mse f1 模型分类 判别生成 监督

## 阶段流程

1. 数据样本

搜集数据样本，质量、数量都要，尽可能贴近真实分布。还可以数据增强，文本可以加字减字错排混插，图像旋转噪音，弥补缺失或者删去。

1. 特征工程

降维

归一化可以压缩参数空间吗，减小过拟合，减小一步线性变换步骤，数值量纲无感的树模型不需要。

分类特征摊开独热编码、序号编码，或者构造出计数、条件概率等编码。

缺失严重可删去，少量缺失可补0、平均值，或者先拟合一遍。

构造组合特征可以发掘高阶信息，范围压缩可以增强泛化性能。

强烈时间属性的，使用近期样本预测。

1. 特征表示

Cnn，rnn，全连接、注意力

根据样本之间的联系，使用图结构提取关系。

1. 度量

线性变换 注意力 余弦 交叉熵

1. 任务

对抗任务、多任务、语义。

多分类任务用多个二分类，还是同时多分类。

1. 训练

基线模型 学习率 超参

1. 预测

效率、泛化、线上线下差异

常见的任务有回归任务、分类任务（二分类、多分类），也有编码表示等中间任务。

准确性、稳健性、效率是模型最重要的几个指标。

对于回归任务，常见的评价指标是方差mse、绝对差mae、对数差RMSLE等，对于分类任务，有全率、确率、准确率、F1，AUC等。

对于分类任务，例如二分类，理想的模型是正例等价于阳性，负例等价于阴性。准确率指

分类正确的概率。有时这两种分类正确还要区分，全率指多少正例被识别为阳性，例如排查传染病。确率是阳性中多少是正例，例如确诊病例。2/F1=1/全率+1/确率。

常常因为阈值变动而引起确率、全率变化，ROC曲线描绘了这种波动情况，以正阴性为代价，正阳性的指标上升有多快，ROC曲线的面积AUC衡量全局下相对指标，不因阈值波动，在排序模型中常用。AUC的直观含义是一样本识别为阳性相对于阴性的概率。另一种计算方式是比较每一个正负样本对，正例预测值大于负例预测值的比率。

生成模型给出先验概率，有直接语义，当然更难。判别模型给出后验概率，要求更低。

有监督模型需要构造目标标签。无监督可以自动构造或者不需要。

## 稳健性

先搜集样本、再构造特征、选取泛化强的基线模型，多类模型集成，损失函数加入复杂度乘法防止过拟合，常见的是L1(损失函数加上参数绝对值)，L2（损失函数加上参数平方），迫使参数尽可能小，不至于极端。

样本不平衡，数据量大就对大类别下采样，数据量少就对小类别上采样，或者采样对抗任务，例如SVM，提高小类惩罚权重。

常常将数据集分为训练集、测试集，有时候也有验证集合。