机器学习基础补充

术语及对应解释

数据集划分

- 训练集（Training set）—— 学习样本数据集，通过匹配一些参数来建立一个模型，主要用来训练模型。类比考研前做的解题大全。
- 测试集（Test set）—— 测试训练好的模型的分辨能力。类比 考研之前做的模拟考试。
- 验证集（validation set）—— 对学习出来的模型，调整模型的参数，如在神经网络中选择隐藏单元数。验证集还用来确定网络结构或者控制模型复杂程度的参数。类比 考研。这次真的是一考定终身。

模型拟合程度

- 欠拟合（Underfitting）：模型没有很好地捕捉到数据特征，不能够很好地拟合数据，对训练样本的一般性质尚未学好。类比，光看书不做题觉得自己什么都会了，上了考场才知道自己啥都不会。
- 过拟合（Overfitting）：模型把训练样本学习“太好了”，可能把一些训练样本自身的特性当做了所有潜在样本都有的一般性质，导致泛化能力下降。类比，做课后题全都做对了，超纲题也都认为是考试必考题目，上了考场还是啥都不会。

通俗来说，欠拟合和过拟合都可以用一句话来说，欠拟合就是：“你太天真了！”，过拟合就是：“你想太多了！”。

常见的模型指标

- 正确率 —— 提取出的正确信息条数 / 提取出的信息条数
- 召回率 —— 提取出的正确信息条数 / 样本中的信息条数
- F 值 —— 正确率 * 召回率 * 2 /（正确率 + 召回率）（F值即为正确率和召回率的调和平均值）

举个例子如下：

某池塘有 1400 条鲤鱼，300 只虾，300 只乌龟。现在以捕鲤鱼为目的。撒了一张网，逮住了 700 条鲤鱼，200 只虾， 100 只乌龟。那么这些指标分别如下：

正确率 = 700 / (700 + 200 + 100) = 70%

召回率 = 700 / 1400 = 50%

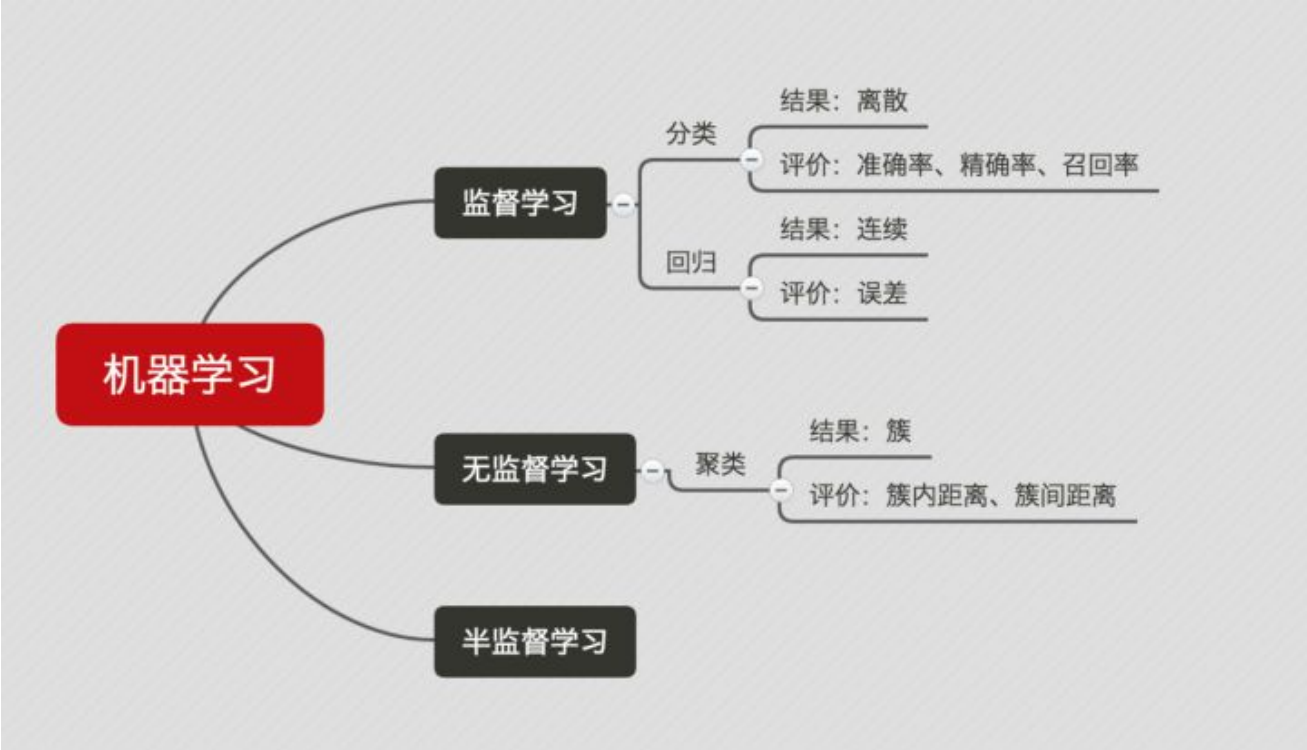
F 值 = 70% * 50% * 2 / (70% + 50%) = 58.3%

模型

- 分类问题 —— 说白了就是将一些未知类别的数据分到现在已知的类别中去。比如，根据你的一些信息，判断你是高富帅，还是穷屌丝。评判分类效果好坏的三个指标就是上面介绍的三个指标：正确率，召回率，F值。
- 回归问题 —— 对数值型连续随机变量进行预测和建模的监督学习算法。回归往往会通过计算 **误差（Error）** 来确定模型的精确性。
- 聚类问题 —— 聚类是一种无监督学习任务，该算法基于数据的内部结构寻找观察样本的自然族群（即集群）。聚类问题的标准一般基于距离：**簇内距离（Intra-cluster Distance）** 和 **簇间距离（Inter-cluster**

Distance）。簇内距离是越小越好，也就是簇内的元素越相似越好；而簇间距离越大越好，也就是说簇间（不同簇）元素越不相同越好。一般的，衡量聚类问题会给出一个结合簇内距离和簇间距离的公式。

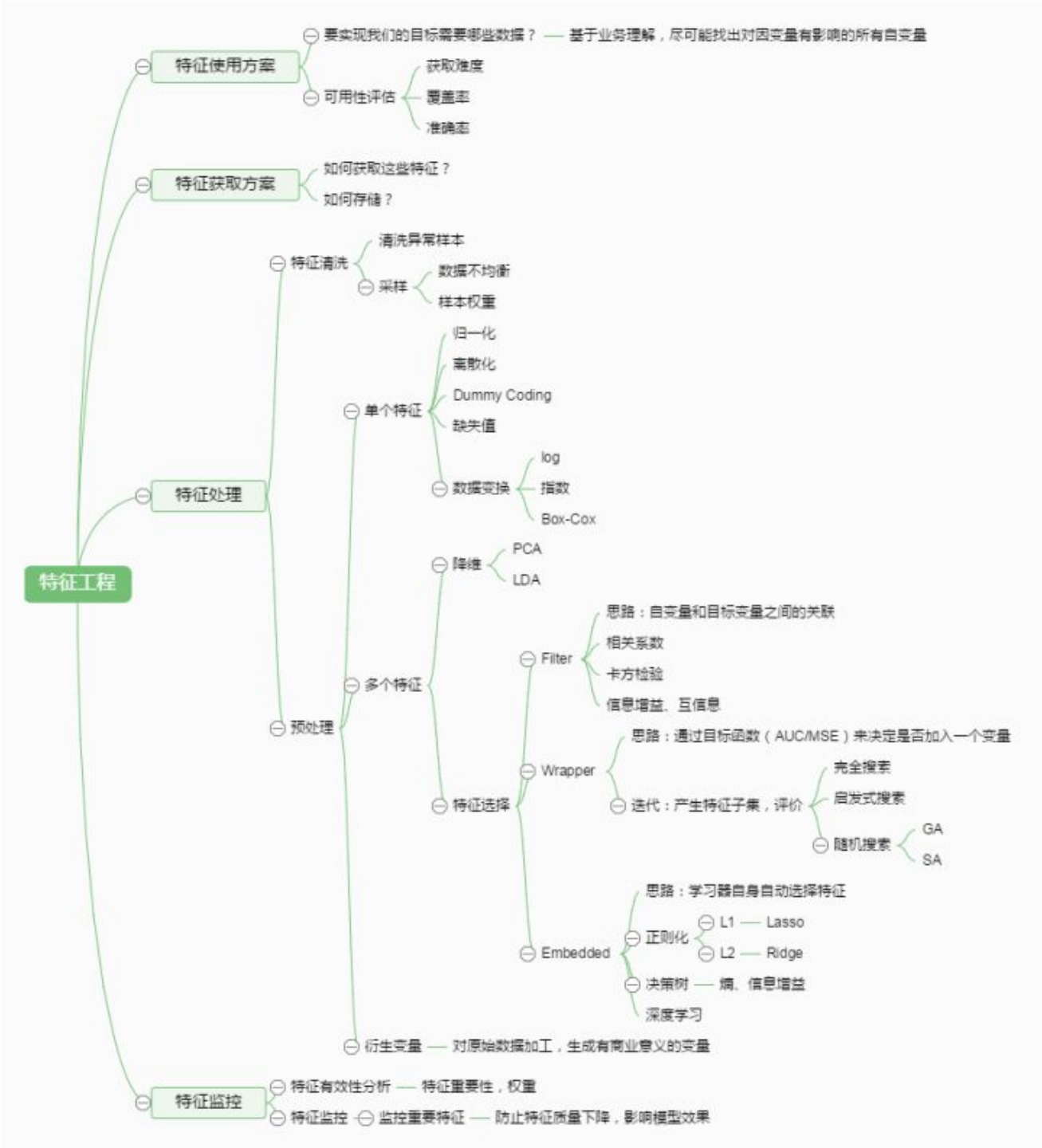
下面这个图比较直观展示出来：



特征工程的一些小东西

- 特征选择 —— 也叫特征子集选择（FSS，Feature Subset Selection）。是指从已有的 M 个特征（Feature）中选择 N 个特征使得系统的特定指标最优化，是从原始特征中选择出一些最有效特征以降低数据集维度的过程，是提高算法性能的一个重要手段，也是模式识别中关键的数据预处理步骤。
- 特征提取 —— 特征提取是计算机视觉和图像处理中的一个概念。它指的是使用计算机提取图像信息，决定每个图像的点是否属于一个图像特征。特征提取的结果是把图像上的点分为不同的子集，这些子集往往属于孤立的点，连续的曲线或者连续的区域。

下面给出一个特征工程的图吧。嘿嘿，偷一下懒~~



其他

- Learning rate —— 学习率，通俗地理解，可以理解成为步长，步子大了，很容易错过最佳结果。就是本来目标尽在咫尺，可是因为我迈的步子很大，却一下子走过了。步子小了呢，就是同样的距离，我却要走很多很多步，这样导致训练的耗时费力还不讨好。
- 有一个总结的知识点超多的链接：https://zhuanlan.zhihu.com/p/25197792