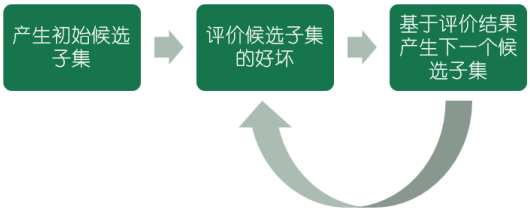
特征工程

特征 = 物体的属性

特征的分类：相关特征（学习任务에 쓸모있는），无关特征(쓸모없는)

特征选择：从给定的特征集合中选出**任务相关特征子集**。确保不丢失重要特征

一般方法：모든 子集를 고려할 수 없으니까



两个关键环节：子集搜索和子集评价

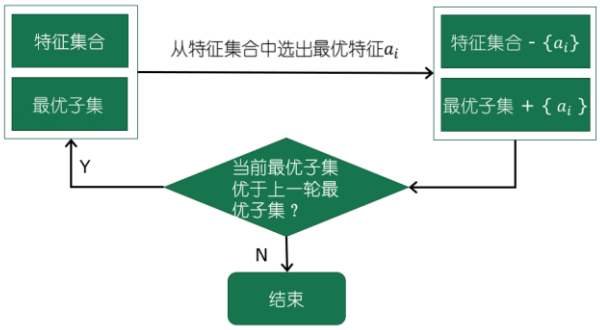
子集搜索：

用贪心策略选择包含重要信息的特征子集

- 前向搜索：逐渐增加相关特征
- 后向搜索：从完整的特征集合开始，逐渐减少特征
- 双向搜索：每一轮逐渐增加相关特征，同时减少无关特征

前向搜索：집합을 2개 만들어서 빼고 더한후, 비교하는 방식으로 순환

- 最优子集初始为空集，特征集合初始时包括所有给定特征



特征子集：确定了对数据集的一个划分

그래서 每个划分区域对应着特征子集的某种取值 / 样本标记对应着对数据集的真实划分

通过估算这两个划分的差异，就能对特征子集进行评价；与样本标记对应的划分的差异越小，则说明当前特征子集越好

그러니까 特征子集的划分이랑 样本标记对应的真实划分을 信息增益로 비교한다

- 特征子集A确定了对数据集D的一个划分
  - A上的取值将数据集D分为V份，每一份用D^v表示
  - $\mathrm{Ent}(D^v)$表示D^v上的信息熵
- 样本标记Y对应着对数据集D的真实划分
  - $\mathrm{Ent}(D)$表示D上的信息熵

$$\mathrm{Ent}(D) = -\sum_{i=1}^{|\mathcal{Y}|} p_k \log_2 p_k ,$$

特征子集A的信息增益为

$$\mathrm{Gain}(A) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v) ,$$

特征选择的常用方法（시험에 나왔음）：

过滤式选择： Relief方法

包裹式选择：LVW方法

嵌入式选择：用L1正则化

过滤式选择

先用特征选择过程过滤原始数据，再用过滤后的特征来训练模型；特征选择过程与后续学习器无关

- Relief (Relevant Features) 方法 [Kira and Rendell, 1992]

是一种用于特征选择的算法，其核心思想是通过评估每个特征对目标分类的区分能力来选择重要的特征。它在处理高维数据、噪声和相关特征时表现良好。

- 核心思想

Relief算法通过随机选取样本并考察其在特征空间中的近邻样本来估计特征的重要性。算法依据的原则是：

- 一个好的特征在同类样本中具有相似值（即类内相似）。
- 一个好的特征在异类样本中具有不同值（即类间差异）。

包裹式选择方法

包裹式选择方法（也称封装式方法）直接把最终将要使用的学习器的性能作为特征子集的评价准则

- 包裹式特征选择的目的就是为给定学习器选择最有利于其性能、“量身定做”的特征子集

- 包裹式选择方法直接针对给定学习器进行优化，因此从最终学习器性能来看，包裹式特征选择比过滤式特征选择更好

- 包裹式特征选择过程中需多次训练学习器，计算开销通常比过滤式特征选择大得多

嵌入式选择

嵌入式特征选择是将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，在学习器训练过程中自动地进行特征选择

- 考虑最简单的线性回归模型，以平方误差为损失函数，并引入L_2范数正则化项防止过拟合，则有

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_2^2$$

岭回归 (ridge regression) [Tikhonov and Arsenin, 1977]

- 将L_2范数替换为L_1范数，则有**LASSO** [Tibshirani, 1996]

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

易获得稀疏解，是一种嵌入式特征选择方法