西瓜书复习笔记11

- **特征选择与稀疏学习：**
  - 什么是特征选择：
  - 子集搜索与评价：
    - 相关特征：
    - 无关特征：
    - 冗余特征：
    - 特征选择：
      - 基于行业基础知识去做
      - 没有行业知识，基于数学角度去做
        迭代法（子集搜索与评价）：
        候选子集-评价-针对评价选下一个子集-直到无法得到更好的候选子集
    - 子集搜索与评价目的：
      从特征中搜索出一个包含所有重要信息的子集
    - 子集搜索：
      前向搜索：
      在最优的子集上逐步增加特征（依据是子集评价），直到增加特征并不能使模型性能提升为止。
      后向搜索：
      所有特征全要，对每个特征进行减少，评价效果是否增加，如果增加了，那就把这个删除。
      双向搜索。
    - 子集评价：
      用信息增益当作指标，信息增益越大意味着该特征包含分类的信息越多。
      对每个候选特征计算其信息增益，选出最大的。
    - 像是特征选择：
      决策树的决策划分其实就可以看作是特征选择，其实是隐式结合了子集搜索机制。
  - 过滤式选择(filter)：
    先进行特征选择，再进行训练。特征选择与训练分开，特征选择与后续学习无关。
    - Relief：
      - 选择方法：
        指定一个阈值 r，然后选择比 r 大的相关统计量分量所对应的特征即可；
        指定要选取的特征个数 k，然后选择相关统计量分量最大的 k 个特征。
      - 如何确定相关统计量：

$$\delta^j = \sum_{i=1}^{n} -\operatorname{diff}\left(x_i^j, x_{i,nh}^j\right)^2 + \operatorname{diff}\left(x_i^j, x_{i,nm}^j\right)^2$$

计算猜对临近和猜错临近。

猜错临近越大说明分类能力越强。利用它计算特征与输出值之间的相关性，相关统计量越大说明特征的分类能力越强。

- 包裹式选择(wrapper)：

  包裹式特征选择直接针对给定学习器进行优化。

  - 优点：从最终学习器的性能来看，包裹式比过滤式更好；
  - 缺点：由于特征选择过程中需要多次训练学习器，因此包裹式特征选择的计算开销通常比过滤式特征选择要大得多。
  - 做法：

    从初始特征集合中不断的选择特征子集，训练学习器，根据学习器的性能来对子集进行评价，直到选择出最佳的子集。（向前搜索，向后搜索，LVW随机产生子集）

- 嵌入式选择：

  把属性权重嵌入到目标函数中，在学习器训练过程中自动地进行特征选择。

  最常用L1、L2做法。

  正则化项越大，模型越简单，权重系数越小。当正则化项增大到一定程度时，所有的特征系数都会趋于0，在这个过程中，会有一部分特征的系数先变成0。也就实现了特征选择过程。