机器学习の特征

- Ⅰ.特征提取
  - 信号表示
    - 描述:The goal of the feature extraction mapping is to represent the samples accurately in a low-dimensional space. 也就是说，特征抽取后的特征要能够精确地表示样本信息，使得信息丢失很小
    - 方法:PCA
  - 信号分类
    - 描述:The goal of the feature extraction mapping is toenhance the class-discriminatory information in a low-dimensional space. 也就是说，特征抽取后的特征，要使得分类后的准确率很高，不能比原来特征进行分类的准确率低
    - 方法:对于线性来说，对应的方法是LDA
- Ⅱ.特征选择
  - a.特征产生
    - 算法分类
      - Complete(完全搜索)
        - BFS(广搜)
          - 描述:枚举了所有的特征组合，属于穷举搜索，时间复杂度是O(2^n)，实用性不高
        - Branch and Bound(分支限界搜索)
          - 描述:在穷举搜索的基础上加入分支限界。例如：若断定某些分支不可能搜索出比当前找到的最优解更优的解，则可以剪掉这些分支。
        - Beam Search(定向搜索)
          - 描述:首先选择N个得分最高的特征作为特征子集，将其加入一个限制最大长度的优先队列，每次从队列中取出得分最高的子集，然后穷举向该子集加入1个特征后产生的所有特征集，将这些特征集加入队列。
        - Best First Search(最优优先搜索)
          - 描述:与定向搜索类似，唯一的不同点是不限制优先队列的长度。
      - Heuristic(启发式搜索)
        - SFS, Sequential Forward Selection(序列前向选择)
          - 描述:特征子集X从空集开始，每次选择一个特征x加入特征子集X，使得特征函数J(X)最优。简单说就是，每次都选择一个使得评价函数的取值达到最优的特征加入，其实就是一种简单的贪心算法。
          - 评价：缺点是只能加入特征而不能去除特征。例如：特征A完全依赖于特征B与C，可以认为如果加入了特征B与C则A就是多余的。假设序列前向选择算法首先将A加入特征集，然后又将B与C加入，那么特征子集中就包含了多余的特征A。
        - SBS, Sequential Backward Selection(序列后向选择)
          - 描述：从特征全集O开始，每次从特征集O中剔除一个特征x，使得剔除特征x后评价函数值达到最优。
          - 评价：序列后向选择与序列前向选择正好相反，它的缺点是特征只能去除不能加入。另外，SFS与SBS都属于贪心算法，容易陷入局部最优值。
        - BDS , Bidirectional Search(双向搜索)
          - 描述：使用序列前向选择(SFS)从空集开始，同时使用序列后向选择(SBS)从全集开始搜索，当两者搜索到一个相同的特征子集C时停止搜索。
        - LRS , Plus-L Minus-R Selection(增L去R选择算法)
          - 类型
            - <1> 算法从空集开始，每轮先加入L个特征，然后从中去除R个特征，使得评价函数值最优。(L > R)
            - <2> 算法从全集开始，每轮先去除R个特征，然后加入L个特征，使得评价函数值最优。(L < R)
          - 评价：增L去R选择算法结合了序列前向选择与序列后向选择思想， L与R的选择是算法的关键。
        - Sequential Floating Selection(序列浮动选择)
          - 描述：序列浮动选择由增L去R选择算法发展而来，该算法与增L去R选择算法的不同之处在于：序列浮动选择的L与R不是固定的，而是“浮动”的，也就是会变化的。
          - 类型
            - SFFS , Sequential Floating Forward Selection(序列浮动前向选择)
              - 描述：从空集开始，每轮在未选择的特征中选择一个子集x，使加入子集x后评价函数达到最优，然后在已选择的特征中选择子集z，使剔除子集z后评价函数达到最优。
            - SFBS , Sequential Floating Backward Selection (序列浮动后向选择)
              - 描述：与SFFS类似，不同之处在于SFBS是从全集开始，每轮先剔除特征，然后加入特征。
              - 评价：序列浮动选择结合了序列前向选择、序列后向选择、增L去R选择的特点，并弥补了它们的缺点。
        - DTM, Decision Tree Method(决策树)
          - 描述:在训练样本集上运行C4.5或其他决策树生成算法，待决策树充分生长后，再在树上运行剪枝算法。则最终决策树各分支处的特征就是选出来的特征子集了。决策树方法一般使用信息增益作为评价函数。
      - Random(随机搜索)
        - RGSS, Random Generation plus Sequential Selection(随机产生序列选择算法)
          - 描述：随机产生一个特征子集，然后在该子集上执行SFS与SBS算法。
          - 评价：可作为SFS与SBS的补充，用于跳出局部最优值。
        - SA, Simulated Annealing(模拟退火算法)
          - 描述:模拟退火算法以一定的概率来接受一个比当前解要差的解，因此有可能会跳出这个局部的最优解，达到全局的最优解。
          - 评价：模拟退火一定程度克服了序列搜索算法容易陷入局部最优值的缺点，但是若最优解的区域太小（如所谓的“高尔夫球洞”地形），则模拟退火难以求解。
        - GA，Genetic Algorithms(遗传算法)
          - 描述：首先随机产生一批特征子集，并用评价函数给这些特征子集评分，然后通过交叉、突变等操作繁殖出下一代的特征子集，并且评分越高的特征子集被选中参加繁殖的概率越高。这样经过N代的繁殖和优胜劣汰后，种群中就可能产生了评价函数值最高的特征子集。
          - 随机算法的共同缺点：依赖于随机因素，有实验结果难以重现。
  - b.评价函数
    - 功能:评价一个特征子集好坏程度的一个准则
    - 类型
      - Filter(筛选器)
        - 描述:与模型无关,基于一些变特征的衡量标准(即给每一个特征打分,表示这个特征的重要程度),排序后除去那些得分较低的特征.
        - 类型
          - Correlation(相关性)
            - 描述:运用相关性来度量特征子集的好坏是基于这样一个假设：好的特征子集所包含的特征应该是与分类的相关度较高（相关度高），而特征之间相关度较低的（冗余度低）。
            - 指标:可以使用线性相关系数(correlation coefficient) 来衡量向量之间线性相关度。
          - Distance Metrics(距离)
            - 描述:运用距离度量进行特征选择是基于这样的假设：好的特征子集应该使得属于同一类的样本距离尽可能小，属于不同类的样本之间的距离尽可能远。
            - 指标:欧氏距离、标准化欧氏距离、马氏距离等
          - Information Gain(信息增益)
            - 描述:信息增益或信息增益率
          - Consistency(一致性)
            - 描述:若样本1与样本2属于不同的分类，但在特征A、 B上的取值完全一样，那么特征子集{A，B}不应该选作最终的特征集.
        - 优点:计算时间上较高效,对于过拟合问题具有较高的鲁棒性
        - 缺点:倾向于选择冗余的特征,因为他们不考虑特征之间的相关性,有可能某一个特征的分类能力很差，但是它和某些其它特征组合起来会得到不错的效果
      - Wrapper(封装器)
        - 描述: 假如有p个特征，那么就会有2p种特征组合，每种组合对应了一个模型。Wrapper类方法的思想是枚举出所有可能的情况，从中选取最好的特征组合。这种方式的问题是：由于每种特征组合都需要训练一次模型，而训练模型的代价实际上是很大的，如果p非常大，那么上述方式显然不具有可操作性。
        - 类型
          - Classifier error rate(错误分类率)
            - 描述:使用特定的分类器，用给定的特征子集对样本集进行分类，用分类的精度来衡量特征子集的好坏。
          - forward search（前向搜索）
            - 描述:初始时假设已选特征的集合为空集，算法采取贪心的方式,逐步扩充该集合，直到该集合的特征数达到一个阈值，该阈值可以预先设定，也可以通过交叉验证获得。
          - backward search（后向搜索）
            - 描述: backward search初始时假设已选特征集合F为特征的全集，算法每次删除一个特征，直到F的特征数达到指定的阈值或者F被删空。该算法在选择删除哪一个特征时和forward search在选择一个特征加入F时是一样的做法。
        - 优点:考虑到特征与特征之间的关联性
        - 缺点:1.当观测数据较少时容易过拟合;2.当特征数量较多时,计算时间增长.
      - Embeded(集成方法)
        - 描述:旨在集合filter和wrapper方法的优点(时间复杂度较低，并且也考虑特征之间的组合关系)
        - 类型
          - 正则化
            - 描述:可以见“简单易学的机器学习算法——岭回归(Ridge Regression)”，岭回归就是在基本线性回归的过程中加入了正则项。我们知道L1正则化自带特征选择的功能，它倾向于留下相关特征而删除无关特征。比如在文本分类中，我们不再需要进行显示的特征选择这一步，而是直接将所有特征扔进带有L1正则化的模型里，由模型的训练过程来进行特征的选择。
        - 优点:集合了前面两种方法的优点
        - 缺点:必须事先知道什么是好的选择
  - c.停止准则
    - 功能:与评价函数相关的，一般是一个阈值，当评价函数值达到这个阈值后就可停止搜索
  - d.验证过程
    - 功能:在验证数据集上验证选出来的特征子集的有效性。