



* 常规 ML frame

数据清洗 → 特征工程 → 分类器

* auto 过程:

A. 自学样本数据

无学习  (特征, 样本数量, 数据类型, 数据统计信息...)
模型推荐. 建议 (自带初始化参数)

B. 自动超参优化 (之前还有特征优化过程)
贝叶斯优化  数据清洗

C. 集成模型合并

* 分类器

1. tree-based model

2. KNN

3. Naive bayes

4. SVM

* 回归器.

1. tree-based model

2. KNN

3. Lasso / Ridge / SVM

* 数据预处理过程.

A. balance B 缺失处理 C 分桶 D 归一化 E 标准化

* 特征筛选

A. PCA. 主成份分析 因子分析

B. tree-based model. logistic Regression weight.

C. 方差. 分位数筛选. LDA 空间投影.

* meta-learning 元学习?

1. 对小数据集进行集合上的估计

1. 数据集 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100

(数据重, 特征重, 特征组合的筛选,
统计属性 ...)

2. 通过贝叶斯优化获取这些小数数据集上的最优初始值.
3. 计算新数据集和已有最优结果的数据特征进行相似选择
4. 选择 top k 的最优相似模型

* 自动超参选择:

1. Grid Search
 2. Random Search
 3. 贝叶斯优化
 4. random forest
- (SMAC)