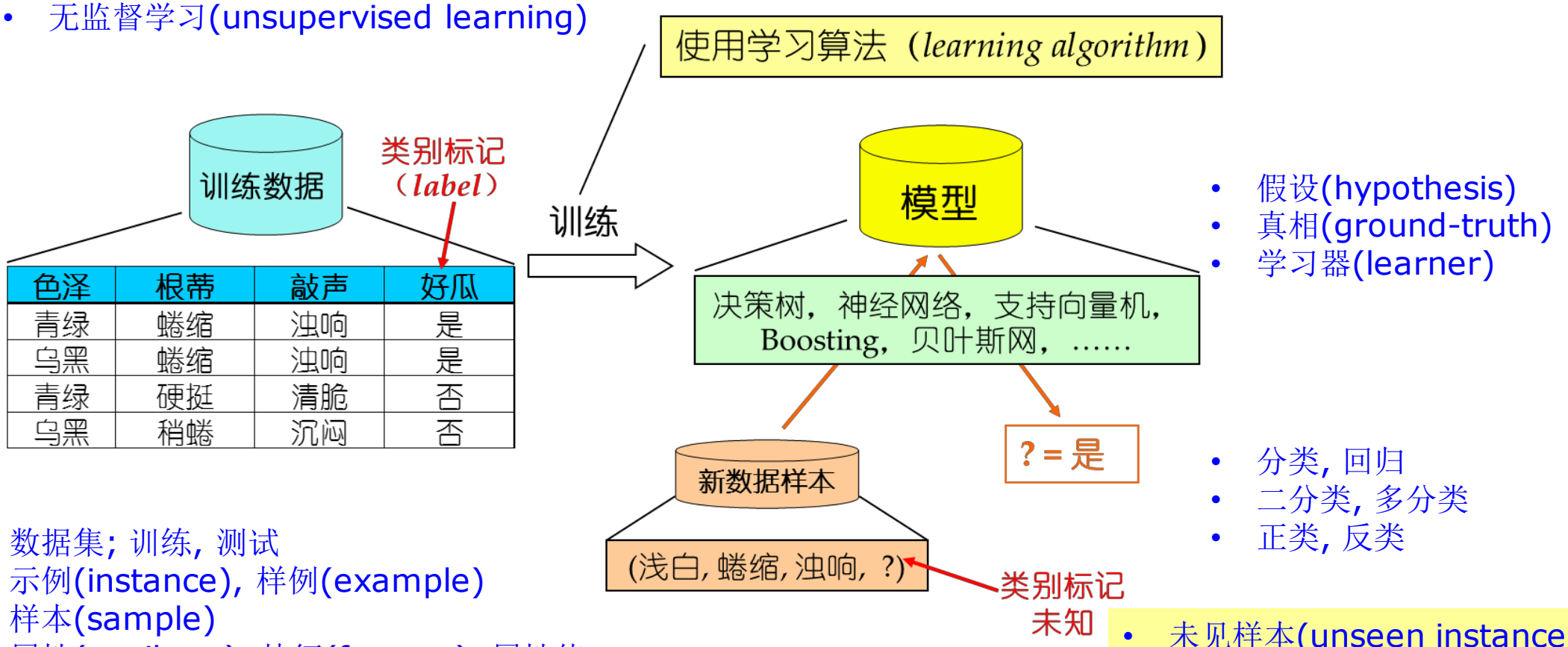
Spark机器学习

寒小阳

2018-01-13

基本术语与概念

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

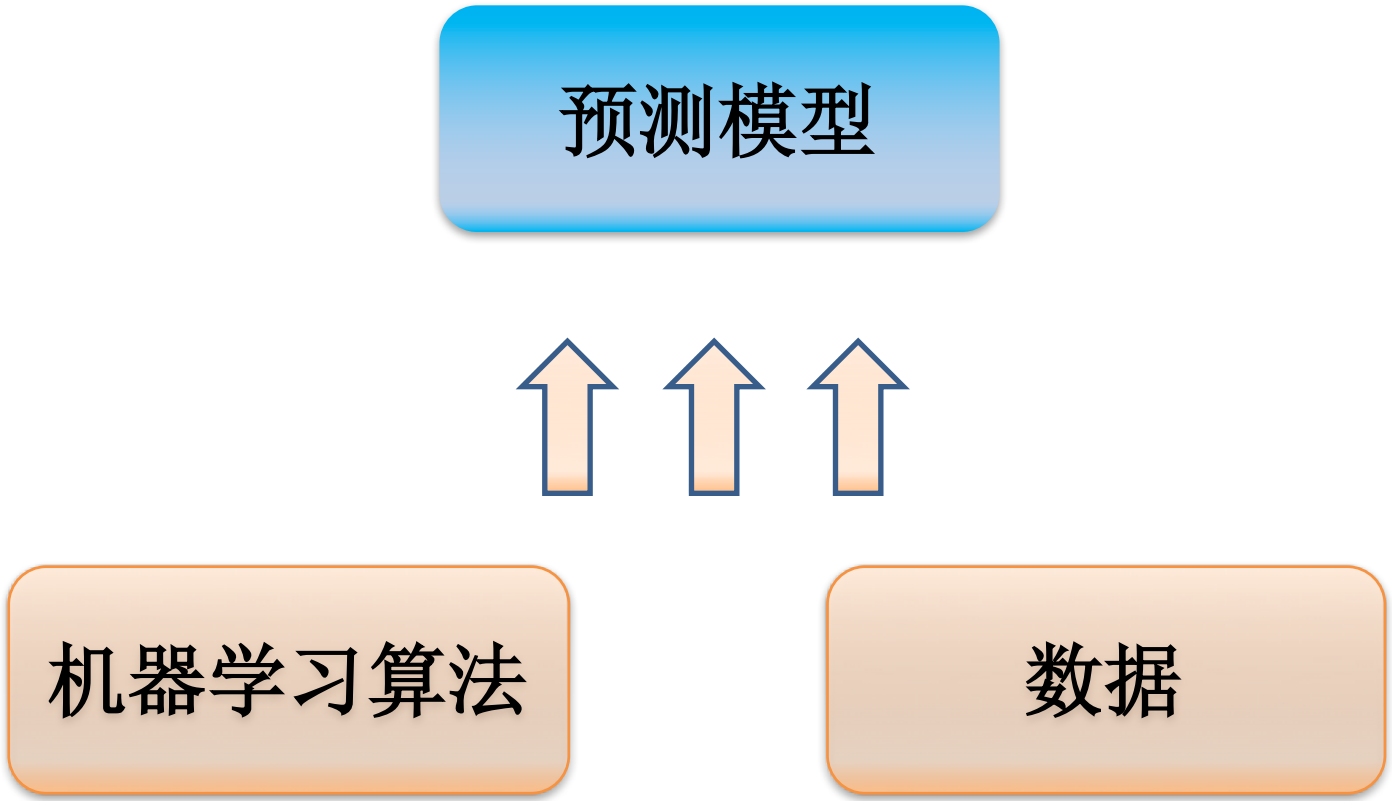
- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类, 回归
- 二分类, 多分类
- 正类, 反类

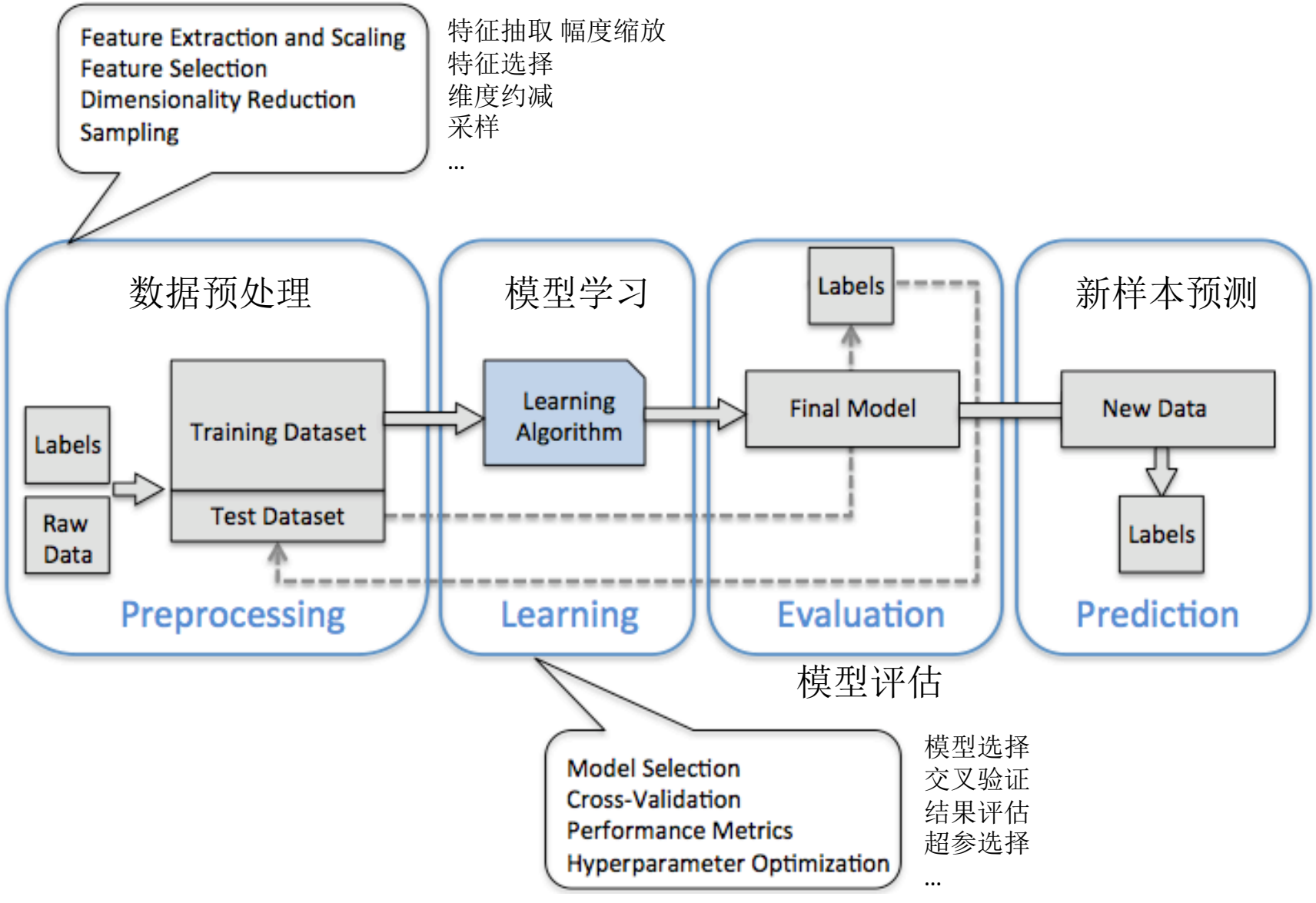
- 数据集; 训练, 测试
- 示例(instance), 样例(example)
- 样本(sample)
- 属性(attribute), 特征(feature); 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

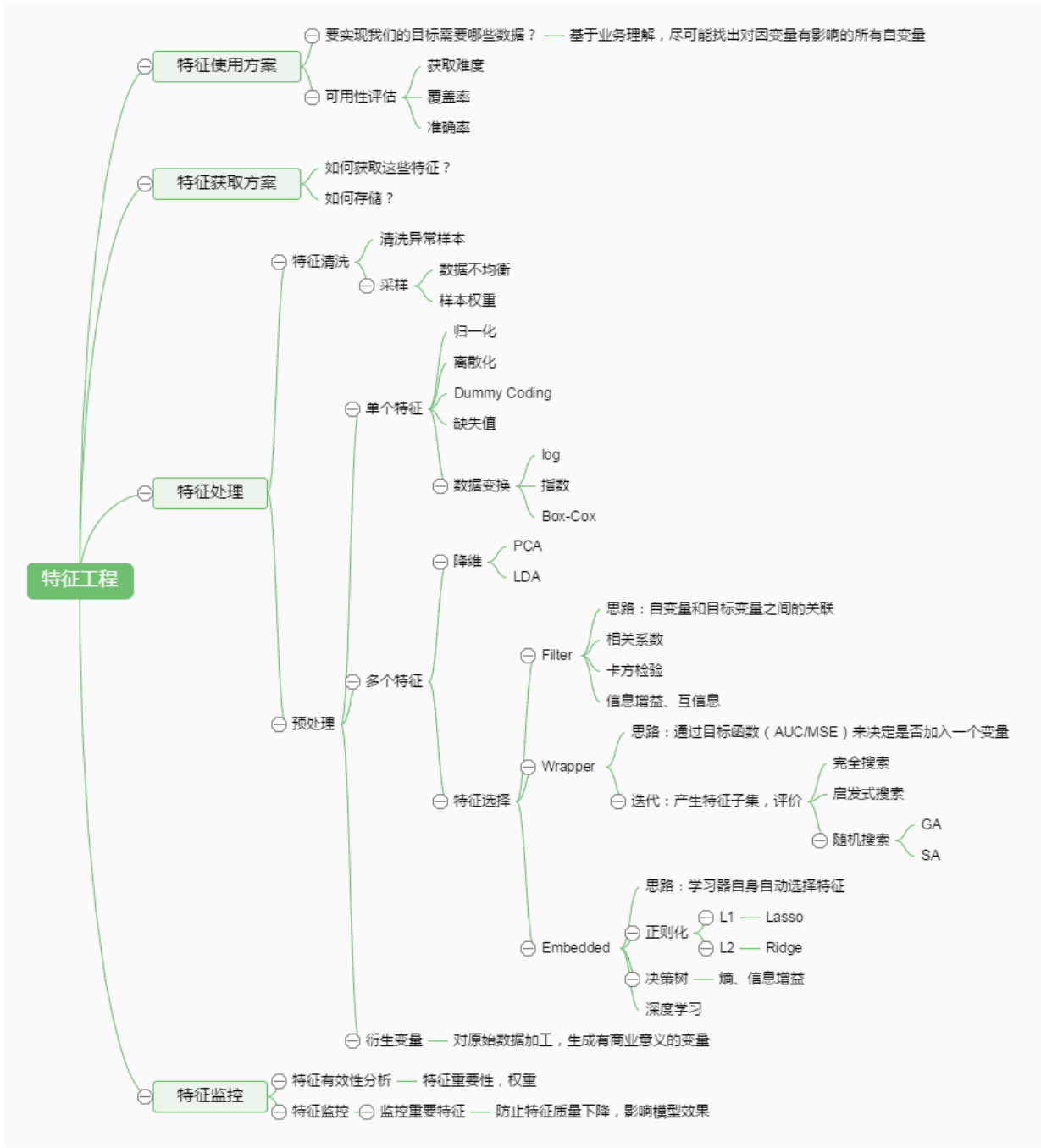
- 未见样本(unseen instance)
- 未知“分布”
- 独立同分布(i.i.d.)
- **泛化(generalization)**

案例 From 周志华《机器学习》

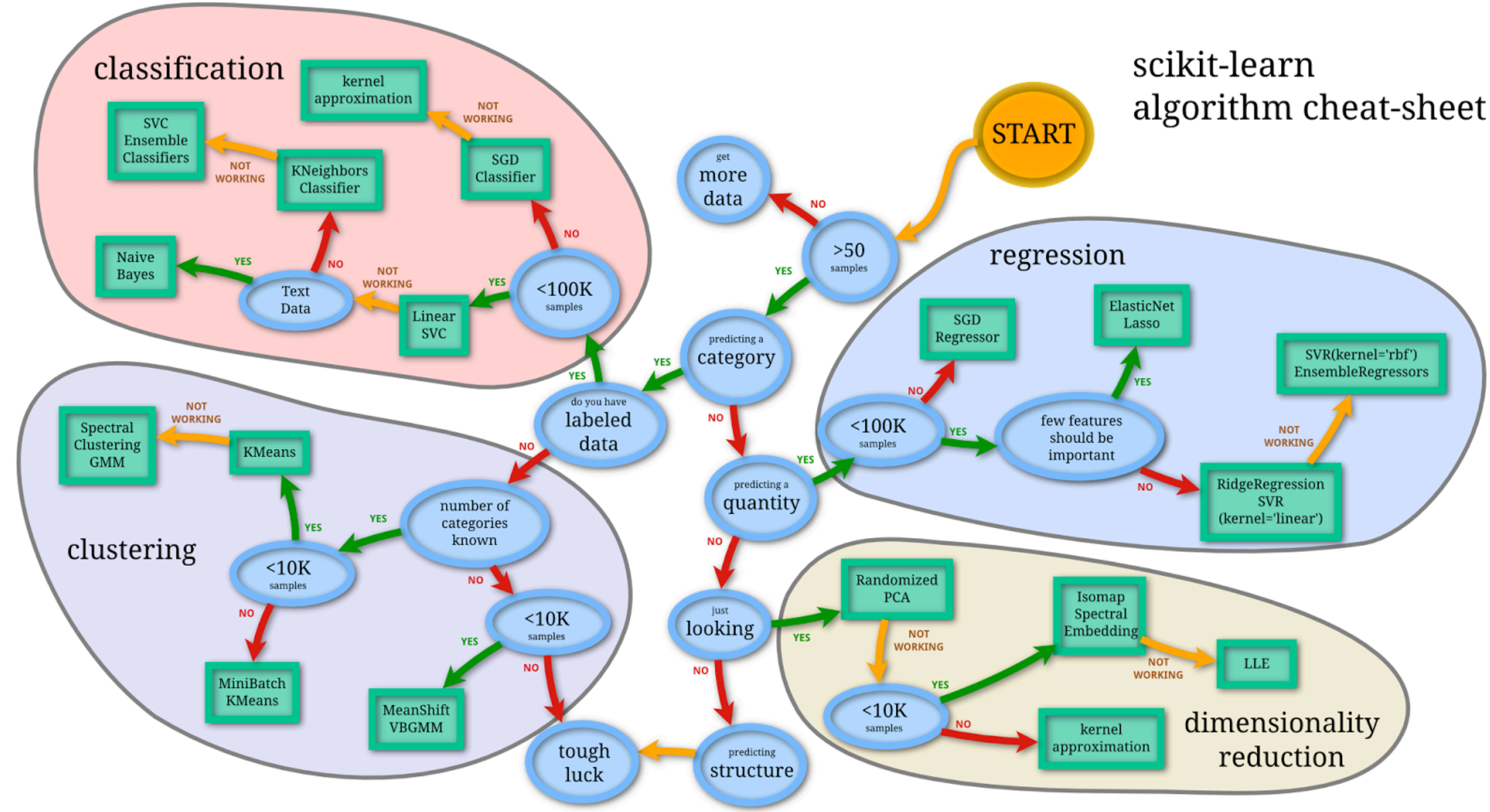


机器学习的应用工作是围绕着数据与算法展开的


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
特征抽取 幅度缩放
特征选择
维度约减
采样
...
数据预处理
模型学习
新样本预测
Labels
Training Dataset
Learning Algorithm
Final Model
New Data
Raw Data
Test Dataset
Labels
Preprocessing
Learning
Evaluation
Prediction
模型评估
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization
模型选择
交叉验证
结果评估
超参选择
...



样本调权
归一化
离散化
独热向量编码
Log/exp变换
PCA


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
NO
predicting a quantity
YES
just looking
NO
predicting structure
tough luck
do you have labeled data
YES
NO
<100K samples
YES
NO
Naive Bayes
Text Data
NOT WORKING
Linear SVC
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
kernel approximation
SGD Classifier
classification
number of categories known
YES
NO
<10K samples
YES
NO
Spectral Clustering GMM
NOT WORKING
KMeans
MiniBatch KMeans
<10K samples
YES
NO
MeanShift VBGMM
clustering
<100K samples
YES
NO
few features should be important
YES
NO
SGD Regressor
ElasticNet Lasso
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
Randomized PCA
YES
NOT WORKING
<10K samples
YES
NO
Isomap Spectral Embedding
NOT WORKING
LLE
kernel approximation
dimensionality reduction

感谢大家！

恳请大家批评指正！