1、评价指标

（1）K折交叉验证（模型参数选择）

---》GridSearchCV函数，可选择超参数的同时做一下交叉验证

---》StratifiedKFold()，cross\_val\_score()只做交叉验证

---》交叉验证用于模型参数选择，评估模型的预测性能，尤其是训练好的模型在新数据上的表现，可以在一定程度上减小过拟合。

---》可以比较出最适合的超参数，如：学习率等

（2）混淆矩阵（评估分类）

---》行为实际类别，列为预测类别，对角线为预测对的个数

（3）准确率和召回率（评估分类）

---》TP：实际为P，预测为P；

---》FP：实际为N，预测为P；

---》TN：实际为N，预测为N；

---》FN：实际为P，预测为N；

准确率(Accuracy) ＝ (TP + TN) / 总样本

精确率(Precision) ＝ TP / (TP + FP)

召回率(Recall) ＝ TP / (TP + FN) 召唤率

（4）F1-Score（F-Measure） = 2/(1/Precision + 1/Recall)

（5）AUC面积（Area under curve）和ROC曲线

---》True Positive Rate( TPR )  = TP / [ TP + FN] ，TPR代表能将正例分对的概率

---》False Positive Rate( FPR ) = FP / [ FP + TN]

---》threshold = 0，全正，（1,1）；threshold = 1，全负，（0,0）

---》model = 100%，完美模型，曲线过（0,1）点，AUC面积为1

（6）决策边界和信心值

===========================================

2、相似度测量（聚类、分类）

（1）聚类和降维

---》把N个维度，通过聚类分为x类降低为x个维度

（2）距离相似度d

---》欧式距离

---》闵可夫斯基距离、曼哈顿距离、切比雪夫距离

（3）jaccard相似系数

---》J(A,B) = (A∩B)/(A∪B)

---》作用：网页去重、查论文、防考试作弊

---》d = 1 - J(A,B)

（4）余弦相似度

---》常用于计算文本相似度

（5）Person相关系数、交叉熵和相对熵

3、聚类

（1）K-Means

---》随机初始中心点

---》计算每点和初始中心点的距离，并划归为最近点的一类

---》求得每一类中的平均点（点坐标均值），为新的中心点

---》反复这一过程，直至没有任何点发生变化，迭代收敛

（2）其他方法

---》K-Mediods：中位数，对K-Mediods的改进，可以减少噪声

---》二分K-Means：用于修正初值选错带来的影响

---》K-Means++：优化初始点的选择（随机选择第一个样本，计算每个样本到第一个样本的距离，距离换算成概率，然后概率化选择）

（3）误差不可能最小

---》对于聚类，每个类别的的误差都为MSE

---》y为每个点的位置，y^为每个类的中心点

（4）Mini Batch K-Means

（5）Canopy聚类

---》与K-Means区别：均匀

---》K不指定

---》点可能同属于多个类别

---》先一次迭代，得出合理K值，再K-Means

================================

4、聚类评估指标

（1）Given Label

---》均一性、完整性

---》V-Measure、F-Measure

（2）轮廓系数

（3）ARI评估

（4）AMI评估

5、层次聚类

---》分裂

---》凝聚（几种方式）

6、密度聚类

---》半径、最小个数

7、谱和谱聚类

---》综合效果最好