# 1.如何把pulearn问题/半监督 转换为监督学习

换句话：如何把一部分label=1的数据集和另一部分label=None的数据集转换为 label 属于{0，1}的数据集

## 方法4.Positive-Unlabel-learning two step（使用聚类方法初始化solid negative sample）

Step 1. identifying reliable negative examples

1.1计算label=1的聚类中心

1.2计算所有点到label=1的聚类中心的距离

1.3选出距离label=1聚类中心最远的那些点 //个数未定，**1000**？？

1.4把这些距离label=1聚类中心最远的点标记为label=0，这些点就是初始的reliable negative examples

Step 2: training a classifier （SVM）

2.1对已经有label的数据（label=1和reliable negative examples）使用classfier\_i训练，把得到的classfier\_i应用到没有label的数据，**找出新的reliable negative examples（只找reliable negative examples不用**classfier\_i找label=1的sample**）**

2.2反复迭代步骤2.1直到：

2.2.1达到最大迭代次数

2.2.2 model的w不再变化 //如果是树，没有w，那么就是数据点的label不再变化

2.3将最终得到的classifier\_n应用到evalset上

## 方法4-2变种.Positive-Unlabel-learning two step（使用孤立森林score初始化solid negative sample）

Step 1. identifying reliable negative examples

1.1将未分类的数据全部默认为0

1.2使用一个模型训练出classifer

1.3用得到的classifier对未标记数据进行predict\_prob

1.4 对未标记数据的predict\_prob排序，选出predict\_prob最小的init\_num个作为solid negative

Step 2: training a classifier（SVM）

2.1对已经有label的数据（label=1和reliable negative examples）使用classfier\_i训练，把得到的classfier\_i应用到没有label的数据，**找出新的reliable negative examples（只找reliable negative examples不用**classfier\_i找label=1的sample**）**

2.2反复迭代步骤2.1直到：

2.2.1达到最大迭代次数

2.2.2 model的w不再变化 //如果是树，没有w，那么就是数据点的label不再变化

2.3将最终得到的classifier\_n应用到evalset上

# 2.如何应对样本不均匀问题

1.对少的样本过采样

2.对多的样本下采样

# 3.模型选择

Model1 GBDT

Model2 Bagging SVM

Model3 RandomForest

Model4 OneclassSVM

Model5 BaggingOneClassSVM

Model6 IsolationForest