**pu learning**

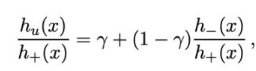
**A bagging SVM to learn from positive and unlabeled examples**

假设unlabeled data 是从正例和负例分布中产生，使用 γ 和 1-γ，挑选正例密度函数为h+,负例密度函数为h-；



增大正例对unlabeled的比对数，同时也能增大正例对负例的比对数，所以能够区分正例和未标记样本的模型理论上也能适用于区分正例和负例。

在实践中发现增大对错误负例的惩罚力度，降低对错误正例的惩罚力度效果会更好。

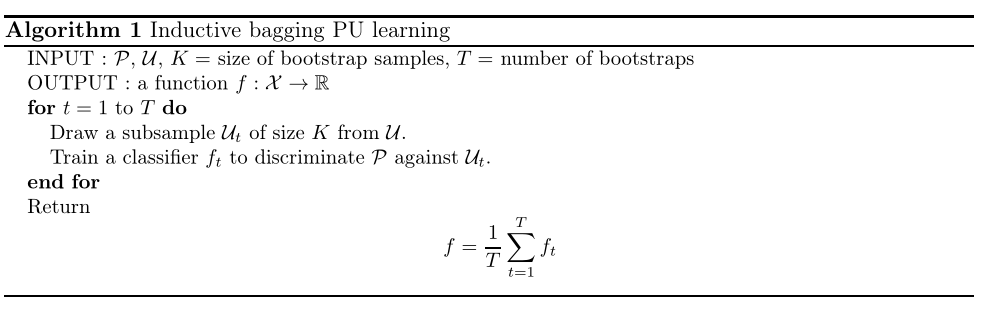


训练过程：

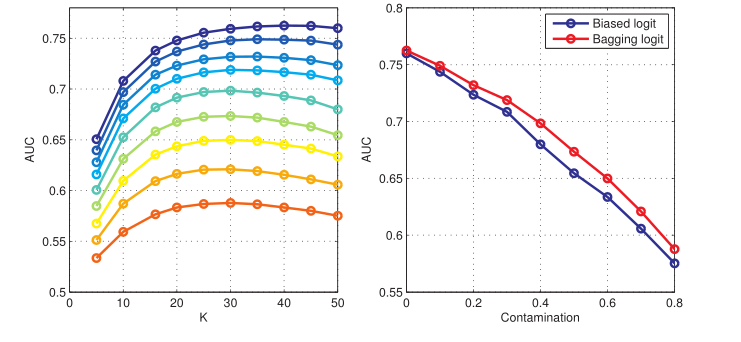
a)通过将所有正样本和未标记样本进行随机组合来创建训练集；

b)利用这个“bootstrap”样本来构建分类器，分别将正样本和未标记样本视为positive和negative；

c)将分类器应用于不在训练集中的未标记样本 - OOB（“out of bag”）- 并记录其分数；

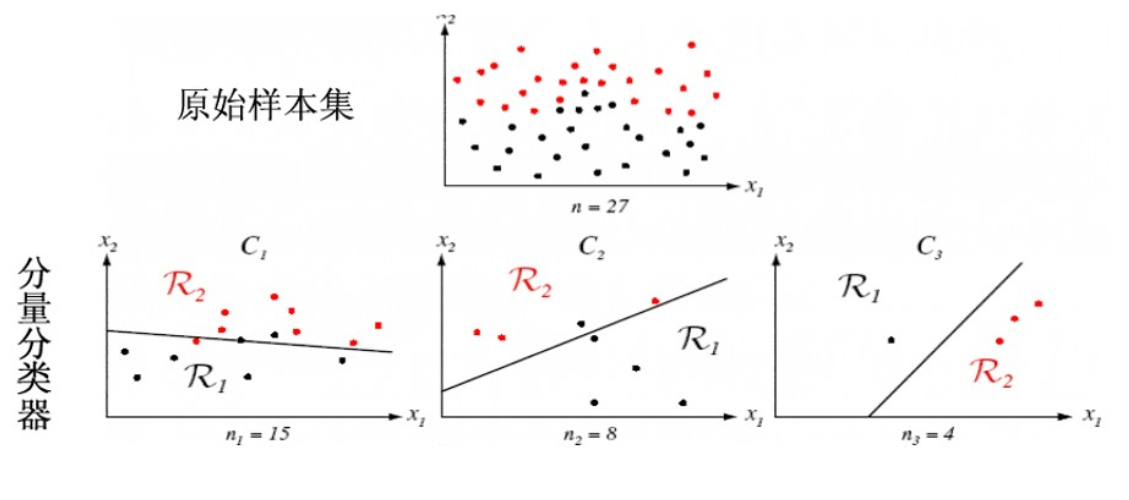
d)重复上述三个步骤，最后为每个样本的分数为OOB分数的平均值。

实验：



从上到下的曲线，依次代表正例样本在未标记样本中占的比例：，可以看出在未标记样本中，正例的个数越多，PU Bagging的效果越差。

问题： 1) Bagging能不能不取均值去学习，考虑优秀的学习器；



· 1)采用哪种策略；

2)考虑在未标记样本集中加入绝对正例；

3) Bagging能不能不取均值去学习，考虑优秀的学习器；

4) 考虑一下分类器模型，决策树系列。

## PU learning+灵活就业人员专利

上周问题：策略、学习模型、结合多方的角度（+，？）、

**策略**：两步法+PU Bagging

**学习模型**：逻辑回归，决策树

**结合多方的角度**：公积金方 A方（+，？），人力资源局 B方（？）

+：正例(positive) ？：未标记样本(unlabeled)

S1：对灵活就业人员多方数据源进行预处理，建立基于多方的灵活就业人员特征数据集；

S2：对A、B方数据源进行样本对齐，将A方的正例样本作为正例，未标记样本作为负例，构建多方联邦二分类模型。

S3：使用该模型进行训练，对双方样本进行筛选，找出可信度较高的负例。

S4：进行二分类，随机抽选已对齐的样本，训练多个联邦基础分类器，根据随机抽选到的绝对负例样本占未标记样本中的比例，调整设置联邦基础分类器相应的权重。

S5：将分类器应用于不在当前随机选择的训练集中的未标记样本 - OOB（“out of bag”）- 并记录其分数。最后每个样本的得分为加权平均后的分数的。根据分数来得到灵活就业推荐人员。

进一步，所述步骤S1中多方数据源指关于灵活就业人员的税务数据、房管数据、社保数据和个人基本信息数据。

使用灵活就业人员的税务数据源，获取税务数据，包括个人所得税。

使用个人基本信息数据源，获取个人基本信息数据，包括。（实例中说明，这里只需要说明使用哪些变量）

进一步地，所述S1中预处理操作包括离群值处理、缺失值处理和数据标准化。

所述离群值处理具体包括：采用箱型图的方式对数据进行可视化统计，剔除其中明显偏离大部分数据的异常值。

所述缺失值处理具体包括：对数据中的每个特征进行缺失率统计，对于缺失数据数目大于总体样本数据规模一半的数据视为无效数据进行剔除，其余数据根据数据分布采用特定方法进行填充，包括众数填充、近邻填充、均值填充和中位数填充。

所述数据标准化具体包括：将已有特征按照数据类型分为连续型特征和离散型特征，对连续型特征采用数据最大-最小归一化处理，对离散型特征采用独热编码处理。

再根据灵活就业人员的公积金缴存情况，将多方源数据分为已缴存灵活就业人员多方源特征数据集和未缴存灵活就业人员多方源特征数据集。其中未缴存灵活就业人员可能包含潜推荐缴存用户。