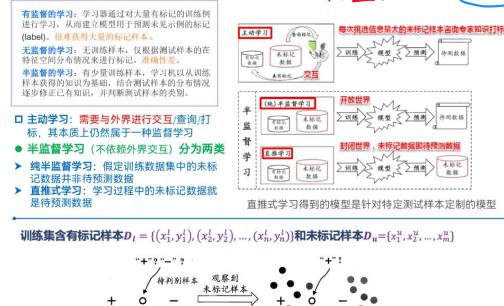


半监督学习：收集数据容易，打标签不容易

泛化能力

在有标签样本较少时如何利用无标签样本提升学习性能



模型在拟合训练上不一致 基于边的方法

- 多视图数据：一个数据对象有多个属性集，每个属性集构成了视图。
- 样本： (x^1, x^2, \dots, y) ，其中“ x ”为样本在视图中的示例， y 为标记。
- 例如电影中的声音和视频分别对应一个视图，类型则为“动作片”、“爱情片”等。

多视图具有兼容性，进而具有互补性

- 不同视图输出空间是一致的，以电影为例，类别均应为（爱情片、动作片）。
- 故可利用多视图的互补性加强分类的准确性。

1 协同训练：基于两个充分且条件独立的视图，利用未标记数据相互促进。

充分：每个视图均包含产生最多学习器的信息。

条件独立：给定类别标记条件下，两个视图独立。

2 基于类标记训练图模型

予以模型挑选出该视图确定的未标记样本簇，并将以上未标记样本作为新的标记样本，从而形成新的训练集。

对视图模型进行训练，进而根据视图模型对未标记样本进行预测，将其加入至视图模型的训练集中。

重复以上两步，直到两个分类器不变。

半监督学习算法的成立依赖于假设：

聚类假设：同一聚类中的样本很可能具有同样的类别标记。关注样本空间的整体特征，探测样本分布稠密和稀疏的区域，从而更好地将决策边界穿过的稀疏区域。

流形假设：高维中的数据存在着低维的特性。利用大量无标签样增加样本空间的密度，从而更准确地获得样本的局部邻域关系，是聚类假设的广。

平滑假设：相似的样本具有相同的标签。

方法概述
生成式方法：样本如何生成 PCC1
半监督模型：样本标签如何 PCC1(a)
GMM 与 SUM

GMM 基础

✓ 已知两个数据分布由两个未知的高斯分布产生，根据观测数据可推断产生这两个高斯分布。

$$\mu_1 = x_1, x_2, \dots, x_n$$

$$\sigma_1^2 = (x_1 - \mu_1)^2 + (x_2 - \mu_1)^2 + \dots + (x_n - \mu_1)^2$$

✓ 数据的分类已知，已知数据由哪两个高斯分布产生，推断样本由哪个高斯分布产生的概率。

$$P(x|C_1) = \frac{P(C_1)P(x|C_1)}{P(C_1)P(x|C_1) + P(C_2)P(x|C_2)}$$

基于 $P(C_1), P(C_2), \mu_1, \mu_2, \Sigma$ ，推断样本属于 C_1 的概率

$$P(C_1|x) = \frac{P(C_1)P(x|C_1)}{P(C_1)P(x|C_1) + P(C_2)P(x|C_2)}$$

通过迭代进行两个步骤，期望步骤（E 步）和最大化步骤（M 步），来估计模型参数。

➢ 有监督的生成模型
• 给定带标签的训练实例 $x \in C_1, C_2$
• 找最可能的先验概率 $P(C_i)$ 和似然概率 $P(x|C_i)$

$P(x|C_1)\mu_1^T \Sigma^{-1} \mu_1 + P(x|C_2)\mu_2^T \Sigma^{-1} \mu_2$

• 基于 $P(C_1), P(C_2), \mu_1, \mu_2, \Sigma$ ，推断样本属于 C_1 的概率

$$P(C_1|x) = \frac{P(C_1)P(x|C_1)}{P(C_1)P(x|C_1) + P(C_2)P(x|C_2)}$$

➢ 半监督的生成模型
• 给定带标签的训练实例 $x \in C_1, C_2$
• 找最可能的先验概率 $P(C_i)$ 和似然概率 $P(x|C_i)$

$P(x|C_1)\mu_1^T \Sigma^{-1} \mu_1 + P(x|C_2)\mu_2^T \Sigma^{-1} \mu_2$

• 无标签数据 x 可以通过该模型重新估计 $P(x|C_1), P(x|C_2)$

参数未知，但样本属于哪一类的信息未知

• 初步假设：减少对轮廓分界线的依赖

Step 1：计算未标记数据的后验概率 $P_{\text{init}}(x|C_1)$

Step 2：更新模型

$$P(C_1|x) = \frac{N_1 \sum_i P(C_1|x^i)}{N_1 \sum_i P(C_1|x^i) + N_2 \sum_i P(C_2|x^i)}$$

$$\mu_1^2 = \frac{1}{N_1} \sum_i x^i \mu_1^T + \frac{1}{N_2} \sum_i x^i \mu_2^T P(C_1|x^i) x^i \dots$$

通过迭代进行两个步骤，期望步骤（E 步）和最大化步

骤（M 步），来估计模型参数

类似生成模型的半监督学习

硬标签 vs. 软标签 [若想使用神经网络]

从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
可对每一个样本赋予一个权重

自监督的可视化：限制分布，使更集中

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集

如何选择移除的数据集
仍是一个问题

• 基于熵的可视化：限制分布，使更集中

• 如何选择移除的数据集

• 根据需要选择合适模型

根据需要选择合适模型

• 将 f* 应用于未标记的数据集

• 得到 $\{(x^u, y^u)\}_{u=1}^R$

• 从未标记数据集中移除一组数据，并将它们添加到带标签数据集