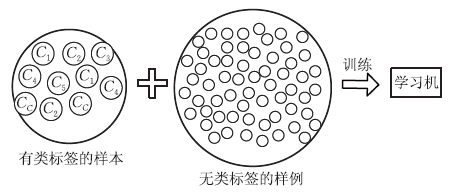
上篇主要介绍了机器学习的理论基础，首先从独立同分布引入泛化误差与经验误差，接着介绍了PAC可学习的基本概念，即以较大的概率学习出与目标概念近似的假设（泛化误差满足预设上限），对于有限假设空间：（1）可分情形时，假设空间都是PAC可学习的，即当样本满足一定的数量之后，总是可以在与训练集一致的假设中找出目标概念的近似；（2）不可分情形时，假设空间都是不可知PAC可学习的，即以较大概率学习出与当前假设空间中泛化误差最小的假设的有效近似（Hoeffding不等式）。对于无限假设空间，通过增长函数与VC维来描述其复杂度，若学习算法满足经验风险最小化原则，则任何VC维有限的假设空间都是（不可知）PAC可学习的，同时也给出了泛化误差界与样本复杂度。稳定性则考察的是输入发生变化时输出的波动，稳定性通过损失函数与假设空间的可学习理论联系在了一起。本篇将讨论一种介于监督与非监督学习之间的学习算法--半监督学习。

#**14、半监督学习**

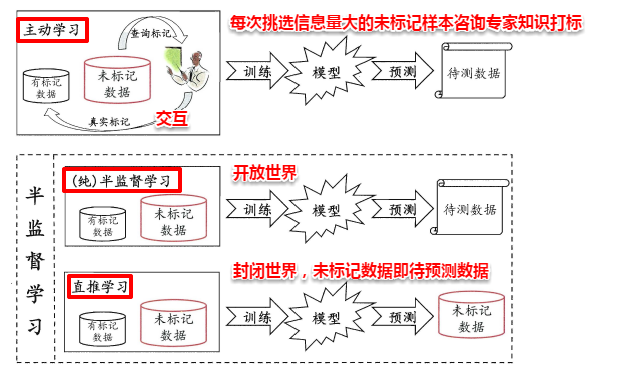
前面我们一直围绕的都是监督学习与无监督学习，监督学习指的是训练样本包含标记信息的学习任务，例如：常见的分类与回归算法；无监督学习则是训练样本不包含标记信息的学习任务，例如：聚类算法。在实际生活中，常常会出现一部分样本有标记和较多样本无标记的情形，例如：做网页推荐时需要让用户标记出感兴趣的网页，但是少有用户愿意花时间来提供标记。若直接丢弃掉无标记样本集，使用传统的监督学习方法，常常会由于训练样本的不充足，使得其刻画总体分布的能力减弱，从而影响了学习器泛化性能。那如何利用未标记的样本数据呢？

一种简单的做法是通过专家知识对这些未标记的样本进行打标，但随之而来的就是巨大的人力耗费。若我们先使用有标记的样本数据集训练出一个学习器，再基于该学习器对未标记的样本进行预测，从中**挑选出不确定性高或分类置信度低的样本来咨询专家并进行打标**，最后使用扩充后的训练集重新训练学习器，这样便能大幅度降低标记成本，这便是**主动学习**（active learning），其目标是**使用尽量少的/有价值的咨询来获得更好的性能**。

显然，**主动学习需要与外界进行交互/查询/打标，其本质上仍然属于一种监督学习**。事实上，无标记样本虽未包含标记信息，但它们与有标记样本一样都是从总体中独立同分布采样得到，因此**它们所包含的数据分布信息对学习器的训练大有裨益**。如何让学习过程不依赖外界的咨询交互，自动利用未标记样本所包含的分布信息的方法便是**半监督学习**（semi-supervised learning），**即训练集同时包含有标记样本数据和未标记样本数据**。

[](https://camo.githubusercontent.com/945602e2871221824776e9f6d3028dda3fa2e3ef/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653339383031642e706e67)

此外，半监督学习还可以进一步划分为**纯半监督学习**和**直推学习**，两者的区别在于：前者假定训练数据集中的未标记数据并非待预测数据，而后者假定学习过程中的未标记数据就是待预测数据。主动学习、纯半监督学习以及直推学习三者的概念如下图所示：

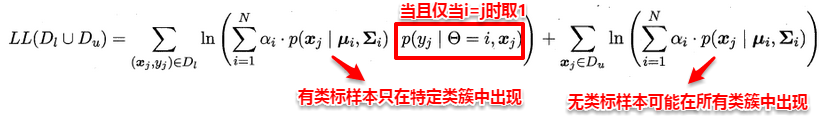
[](https://camo.githubusercontent.com/d84164e056d239c27b9253febfb82dd392b474e6/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a464a532e706e67)

##**14.1 生成式方法**

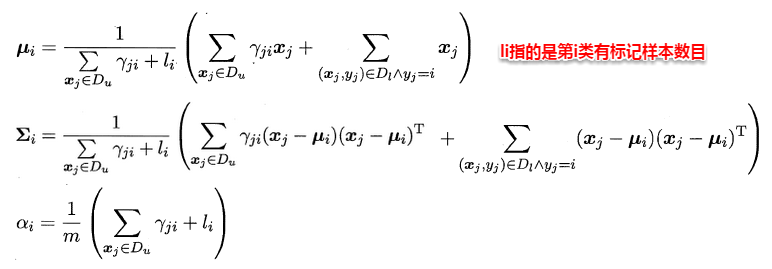
**生成式方法**（generative methods）是基于生成式模型的方法，即先对联合分布P（x,c）建模，从而进一步求解 P（c | x），**此类方法假定样本数据服从一个潜在的分布，因此需要充分可靠的先验知识**。例如：前面已经接触到的贝叶斯分类器与高斯混合聚类，都属于生成式模型。现假定总体是一个高斯混合分布，即由多个高斯分布组合形成，从而一个子高斯分布就代表一个类簇（类别）。高斯混合分布的概率密度函数如下所示：

[](https://camo.githubusercontent.com/9311461e771fd448f48181757b3a3cb22ee538b5/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653362383264632e706e67)

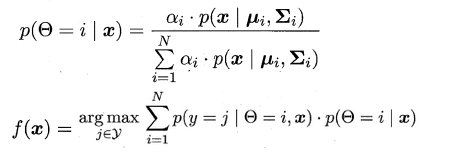
不失一般性，假设类簇与真实的类别按照顺序一一对应，即第i个类簇对应第i个高斯混合成分。与高斯混合聚类类似地，这里的主要任务也是估计出各个高斯混合成分的参数以及混合系数，不同的是：对于有标记样本，不再是可能属于每一个类簇，而是只能属于真实类标对应的特定类簇。

[](https://camo.githubusercontent.com/7cb5c8eb322acd1339f6c4bb065d0ed915e4353e/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653433316433302e706e67)

直观上来看，**基于半监督的高斯混合模型有机地整合了贝叶斯分类器与高斯混合聚类的核心思想**，有效地利用了未标记样本数据隐含的分布信息，从而使得参数的估计更加准确。同样地，这里也要召唤出之前的EM大法进行求解，首先对各个高斯混合成分的参数及混合系数进行随机初始化，计算出各个PM（即γji，第i个样本属于j类，有标记样本则直接属于特定类），再最大化似然函数（即LL（D）分别对α、u和∑求偏导 ），对参数进行迭代更新。

[](https://camo.githubusercontent.com/6aefc06e4739dca177dfd89eb62691d158ad9170/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653433666630382e706e67)

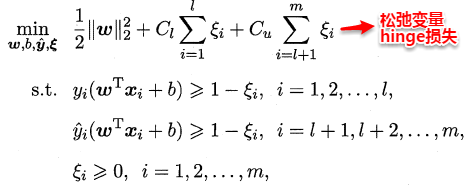
当参数迭代更新收敛后，对于待预测样本x，便可以像贝叶斯分类器那样计算出样本属于每个类簇的后验概率，接着找出概率最大的即可：

[](https://camo.githubusercontent.com/8fb27cfb34cf9f5527f1db7246d323834335de13/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653364666231632e706e67)

可以看出：基于生成式模型的方法十分依赖于对潜在数据分布的假设，即假设的分布要能和真实分布相吻合，否则利用未标记的样本数据反倒会在错误的道路上渐行渐远，从而降低学习器的泛化性能。因此，**此类方法要求极强的领域知识和掐指观天的本领**。

##**14.2 半监督SVM**

监督学习中的SVM试图找到一个划分超平面，使得两侧支持向量之间的间隔最大，即“**最大划分间隔**”思想。对于半监督学习，S3VM则考虑超平面需穿过数据低密度的区域。TSVM是半监督支持向量机中的最著名代表，其核心思想是：尝试为未标记样本找到合适的标记指派，使得超平面划分后的间隔最大化。TSVM采用局部搜索的策略来进行迭代求解，即首先使用有标记样本集训练出一个初始SVM，接着使用该学习器对未标记样本进行打标，这样所有样本都有了标记，并基于这些有标记的样本重新训练SVM，之后再寻找易出错样本不断调整。整个算法流程如下所示：

[](https://camo.githubusercontent.com/44c6d775064dd23baf3b6d0f5a9a983471893a3e/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383536653432373833302e706e67)

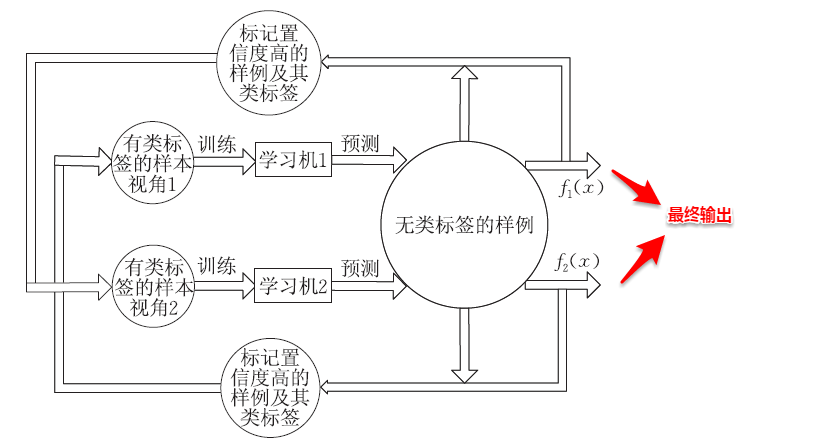
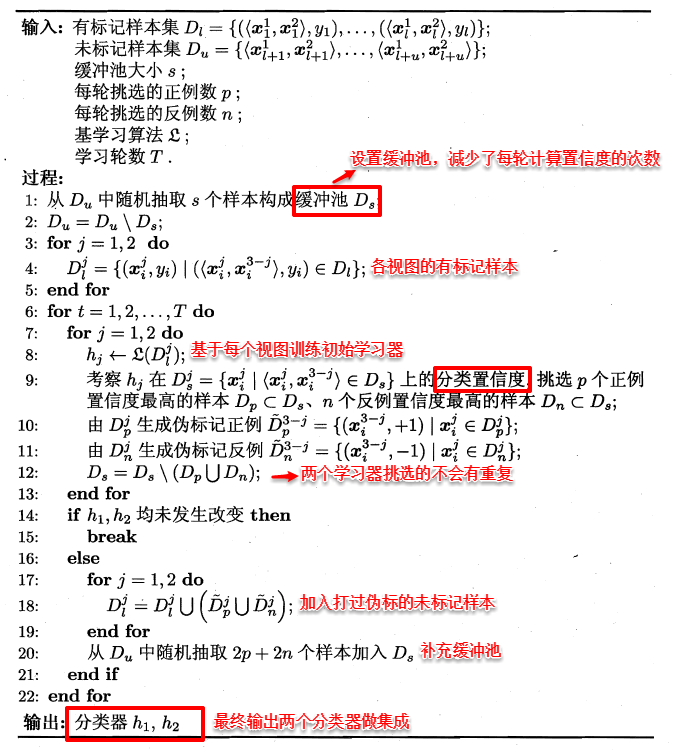
[](https://camo.githubusercontent.com/25c7764bbd397c3e043b7ea36ef340e3014319d2/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a5a73732e706e67)

##**14.3 基于分歧的方法**

基于分歧的方法通过多个学习器之间的\*\*分歧（disagreement）/多样性（diversity）\*\*来利用未标记样本数据，协同训练就是其中的一种经典方法。**协同训练最初是针对于多视图（multi-view）数据而设计的，多视图数据指的是样本对象具有多个属性集，每个属性集则对应一个试图**。例如：电影数据中就包含画面类属性和声音类属性，这样画面类属性的集合就对应着一个视图。首先引入两个关于视图的重要性质：

**相容性**：即使用单个视图数据训练出的学习器的输出空间是一致的。例如都是{好，坏}、{+1,-1}等。 **互补性**：即不同视图所提供的信息是互补/相辅相成的，实质上这里体现的就是集成学习的思想。

协同训练正是很好地利用了多视图数据的“**相容互补性**”，其基本的思想是：首先基于有标记样本数据在每个视图上都训练一个初始分类器，然后让每个分类器去挑选分类置信度最高的样本并赋予标记，并将带有伪标记的样本数据传给另一个分类器去学习，从而**你依我侬/共同进步**。

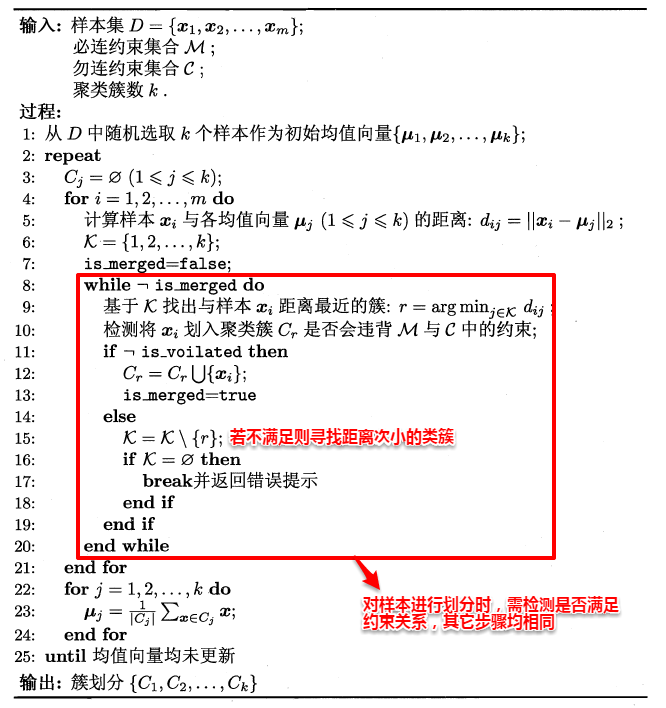
[](https://camo.githubusercontent.com/9fbfa356daaee6f9b012edb8e6ac1445300cf424/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a564d6a2e706e67) [](https://camo.githubusercontent.com/4846d70083bad40b5f9cbf3e30c4067615f5371f/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a654c6e2e706e67)

##**14.4 半监督聚类**

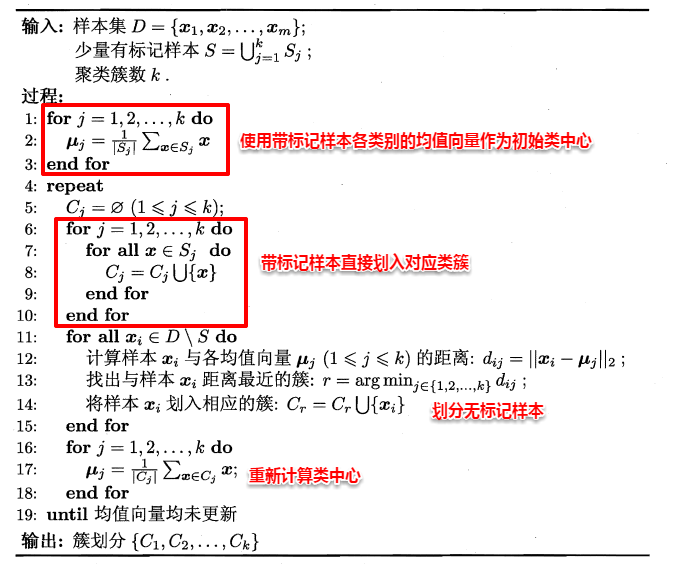
前面提到的几种方法都是借助无标记样本数据来辅助监督学习的训练过程，从而使得学习更加充分/泛化性能得到提升；半监督聚类则是借助已有的监督信息来辅助聚类的过程。一般而言，监督信息大致有两种类型：

**必连与勿连约束**：必连指的是两个样本必须在同一个类簇，勿连则是必不在同一个类簇。 **标记信息**：少量的样本带有真实的标记。

下面主要介绍两种基于半监督的K-Means聚类算法：第一种是数据集包含一些必连与勿连关系，另外一种则是包含少量带有标记的样本。两种算法的基本思想都十分的简单：对于带有约束关系的k-均值算法，在迭代过程中对每个样本划分类簇时，需要**检测当前划分是否满足约束关系**，若不满足则会将该样本划分到距离次小对应的类簇中，再继续检测是否满足约束关系，直到完成所有样本的划分。算法流程如下图所示：

[](https://camo.githubusercontent.com/d6940cd6a6fd53c34e3d9fd4b2a068733a9bf2ae/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a417a512e706e67)

对于带有少量标记样本的k-均值算法，则可以**利用这些有标记样本进行类中心的指定，同时在对样本进行划分时，不需要改变这些有标记样本的簇隶属关系**，直接将其划分到对应类簇即可。算法流程如下所示：

[](https://camo.githubusercontent.com/2ac013f4a52f6e95f531383aab9d6f9f27237fcf/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69774a6b52672e706e67)

在此，半监督学习就介绍完毕。十分有趣的是：半监督学习将前面许多知识模块联系在了一起，足以体现了作者编排的用心。结合本篇的新知识再来回想之前自己做过的一些研究，发现还是蹚了一些浑水，也许越是觉得过去的自己傻，越就是好的兆头吧~