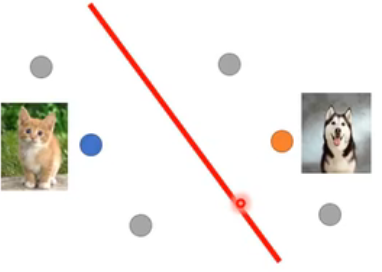
# 机器学习笔记

# Semi-supervised Learning(半监督学习)

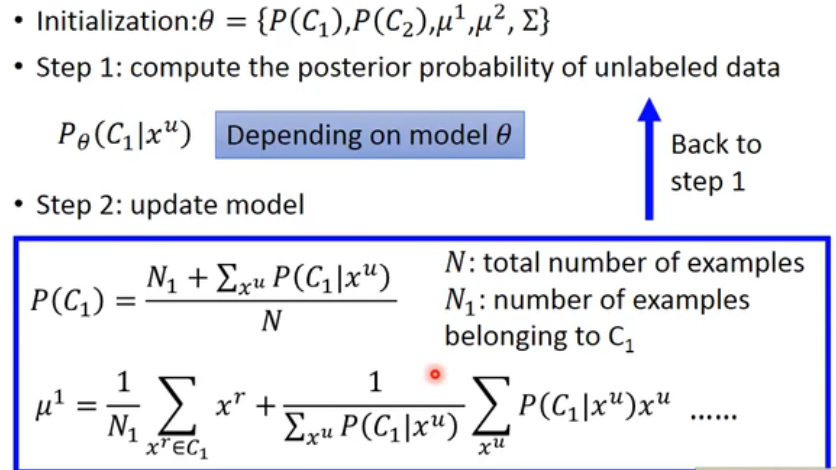
前面学习过的learning方法都是监督式学习，data的输入包括了它的标签，但是在实际上我们手中的data大多数很可能是没有标签的，只有少数的data是拥有标签，这种时候机器就不知道如何用这些没有标签的data来进行训练了，而半监督式学习就可以解决这个问题。

虽然我们不知道一些data的标签，但是我们却可以从它们的分布中得到一些信息。比如下面两幅图，有颜色的点是带标签的data，而灰色的点则是没有标签的data，在左边的图中我们给两种class所划出来的边界是一条竖线，而加入了unlabeled data之后，我们根据自觉可以划出右图所示的斜线作为两种class的边界。但是这个划分好不好取决于我们做的假设符不符合实际，如果灰色点都属于狗的话，那么这个划分就很明显是不好的。



## Semi-supervised Learning for Generative Model

在generative model中，我们用对数据构建高斯模型，然后用高斯模型来求出每笔data属于某个class的几率，然后就可以决定划分class的边界在哪。但是如果我们有unlabeled data的话，就有可能影响高斯模型的形状和分布，在考虑unlabeled data时预测data属于某个class的概率公式如图所示。



每次得到新的model之后我们就重新从步骤一开始再做一次运算。理论上该方法是收敛的，但是收敛的结果会受到初始值的影响。这个方法称为EM算法。

## Low-density Separation

这个做法是基于一个假设：“世界是非黑即白的。”意思是在不同的class之间有很明确的界限，在该界限周围出现data的概率很低。

我们现在手中有labelled data集合，以及unlabeled data集合，接下来我们就重复进行以下步骤：

1.用labelled data训练一个模型.

2.将这个模型应用在unlabeled data set里面.

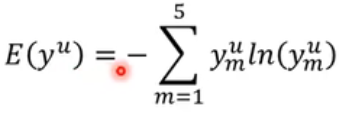
3.从unlabeled data set之中拿一些data放到labelled data set之中.

这个方法不能用于regression上。

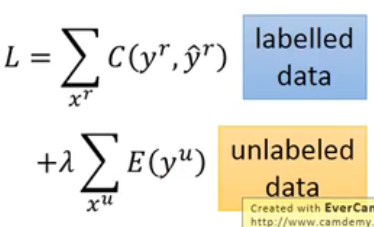
## Entropy-based Regularization

在这个方法中，我们要让output的分布很集中。如下图，前两种的结果是好的，因为output的分布很集中，而最后一种结果是不好的，因为output的分布非常分散。



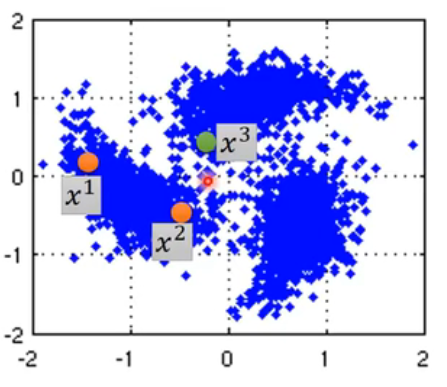
计算分布的熵的公式为，熵越大代表这个分布越差。

根据这个我们可以重新设计loss function，如图所示，loss function的计算既要考虑labelled data，也要考虑unlabeled data。



## Smoothness Assumption

这个做法有一个假设：“近朱者赤，近墨者黑”。即相近的输入会有同样的输出。如果输入x1和x2在同一个高密度区域中距离比较近的话，他们就会对应同样的输出。如下图，x1和x2之间可以通过高密度区域相连，所以x1和x2属于同一个class，而x2和x3之间无法通过一个高密度区域相连，所以x2和x3应该属于不同的class。



这个做法的想法类似演变，如下图最左边的图形既可能像数字“2”，也可能像数字“3”，但是如果这个图形与数字“2”之间有若干个可以“演变”到数字“2”的图形，我们就可以说这个图形是数字“2”。

