姓名：乔翱

学号：202222080625

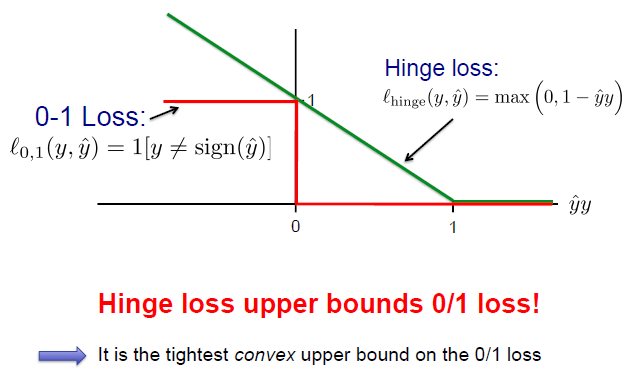
1、支持向量机

以结构风险+经验风险的形式给出支持向量机SVM的目标函数，并说明结构风险和经验风险所起的作用。给出铰链损失（Hinge Loss）函数的形式，以及逻辑回归的对数损失（Log-loss）函数形式，并对它们画图比较。有一个能被正确分类并且远离决策面的样本点，试说明为什么SVM的决策面不会受到它的影响，而逻辑回归的决策面会受到影响。

SVM的目标函数：

经验风险指的是分类器在训练数据上的错误率，而结构风险是对分类器复杂度的度量，惩罚分类器对数据的过拟合，结构风险越小，意味着分类器的复杂度月底，从而降低了泛化误差。

铰链损失函数和逻辑回归的对数损失



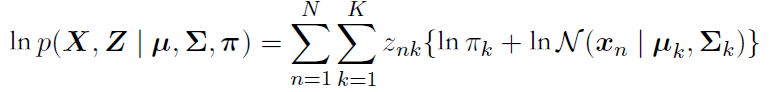
SVM的思想是间隔最大化，对于一个能被正确分类并且远离决策面的样本点，不会对SVM的决策面产生影响，因为其对最大化间隔的目标函数的贡献较小。而逻辑回归使用的是最大似然估计的方法，也就是最大化训练样本的分类概率。其决策面会受到所有训练样本点的影响，包括能被正确分类并且远离决策吗的样本点，该点会导致决策面的偏移。

2、试写出EM算法的流程，写出混合高斯模型完整数据（观测变量与隐变量）的对数似然函数，分别写出混合高斯模型EM算法的E步和M步的计算公式。

EM算法通常用来处理含有隐变量的概率模型参数估计问题。该算法的流程如下：

1. 初始化参数，随机选取参数的初值。
2. E步骤，根据当前参数估计隐变量的后验概率，即计算每个样本在当前参数下属于每个隐变量的概率。
3. M步骤，根据上一步计算得到的后验概率，更新参数估计值，使得对数似然函数最大化。
4. 计算对数似然函数的增量。计算新旧参数估计值之间对数似然函数的差值，如果小于一个预设的阈值，则停止迭代；否则返回E步骤。

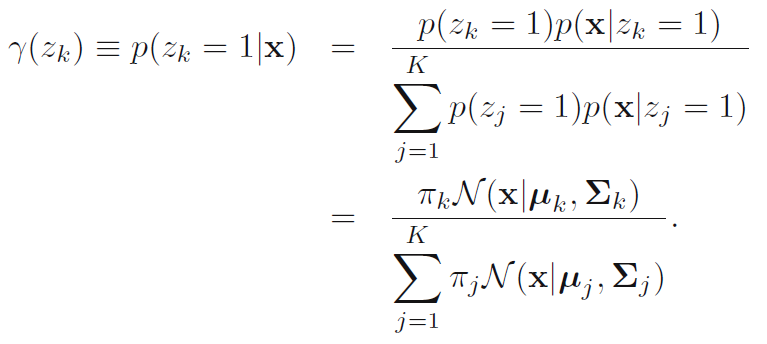
混合高斯模型的完整数据对数似然函数如下：

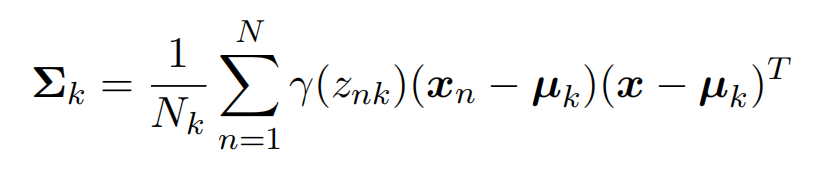


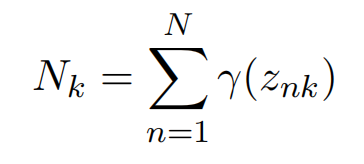
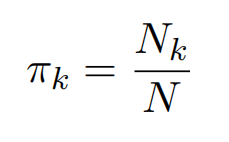
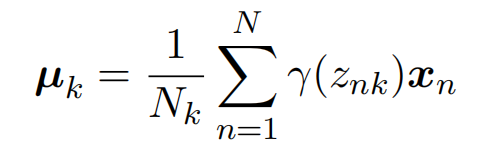
用于混合高斯模型EM描述

1. 为三个参数选择一个初始值
2. 交替进行E步骤和M步骤

E步骤：使用参数的当前值来计算后验概率



M步骤：重新估计



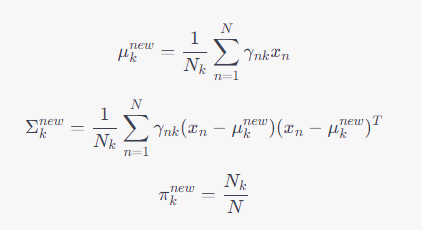
计算对数似然函数，检查参数或者对数似然函数的收敛性，如果没有满足收敛的准则返回E步骤。

1. 用有K个分量的混合高斯模型对数据进行聚类，请问这K个高斯分量代表什么，用EM算法求解时，E步和M步的任务分别是什么，请给出公式和文字说明。如果采用K均值（K-means）算法聚类时，E步和M步的任务分别是什么，请说明两种聚类算法的区别与联系。

K个高斯分量代表k个聚类簇，每个高斯分量表示一个聚类簇的概率密度函数，其均值向量和协方差矩阵分别对应该聚类簇的中心和形状。

E步：给定当前模型参数，计算每个样本点属于每个聚类簇的后验概率。

M步：重新估计参数，最大化对数似然函数来更新高斯分布的混合系数、均值和协方差矩阵。



迭代这两个步骤直到收敛。

使用K均值算法聚类时，E步：计算每个样本点与各个簇中心的距离，并将样本点分配到距离最近的簇中心所对应的簇中。M步：计算每个簇的新中心，即将该簇中所有样本点的坐标取平均值，得到一个新的簇中心。

高斯混合模型的EM算法就像是K均值算法的一个带有固定先验和协方差的软版本。

K均值算法对数据点的聚类进⾏了硬分配，即每个数据点只属于唯⼀的聚类。在K均值算法中，权重为0或1。⽽EM算法基于后验概率分布，进⾏了⼀个软分配。每个中心通过数据的加权方式移动，权重由软分配给定。与K均值算法相⽐，EM算法在达到（近似）收敛之前，经历了更多次的迭代，每次迭代需要更多的计算量。通常运⾏K均值算法找到⾼斯混合模型的⼀个合适的初始化值。协⽅差矩阵可以很⽅便地初始化为通过K均值算法找到的聚类的样本协⽅差。EM不保证找到对数似然函数的全局最大值。