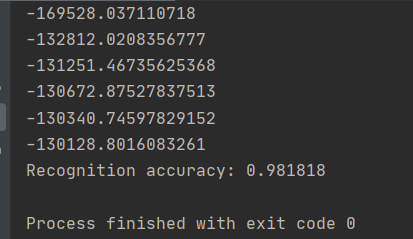
Q：实现基于GMM的0(o)-9孤立词识别系统

提供的数据：本次课程提供了330句训练预料，每个英文单词（0-9，o）含有30个句子用于训练对应的GMM，所有

的训练数据和测试数据的原始音频路径、对应的抄本text（标注，0-9，o）、特征（feats.scp, feats.ark) 都在train和test文件夹下。原始音频的39维MFCC特征已经通过kaldi提取给出，代码中也给出了读取kaldi格式特征的代码。feats.scp 里面存储的是某句话的特征数据的真实文件和位置，特征实际存储在二进制文件feats.ark中（可以忽略kaldi特征部分，我们已经提供了特征读取代码，读取后可在python环境中查看）

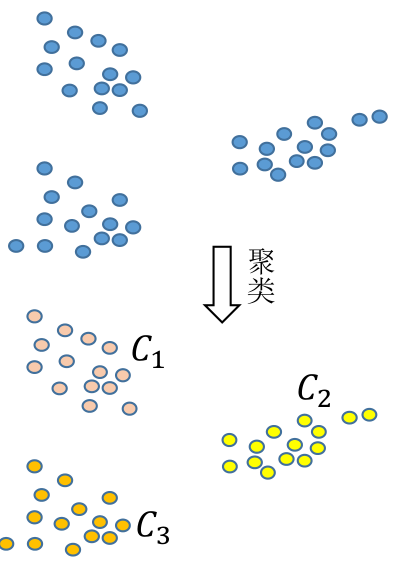
使用提供的特征，完善代码中GMM参数估计部分，并且用测试数据对其进行测试，统计错误率。每一个孤立词建立一个GMM模型，高斯成分个数可以自定，特征维度是39维。

A：识别率98.18%

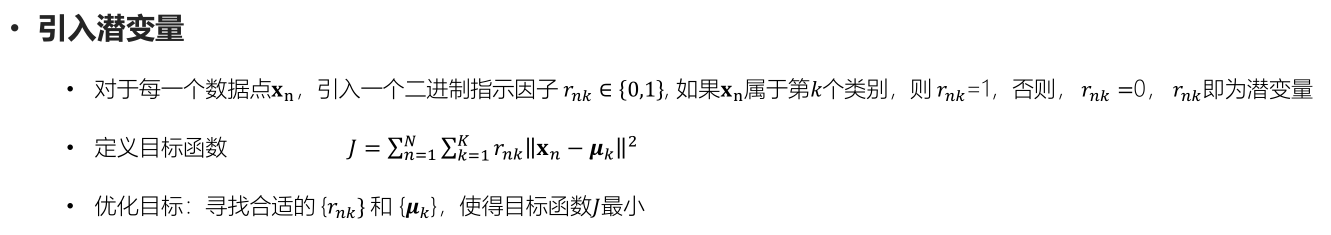


观测变量(observed variable):直接可以观测到的变量

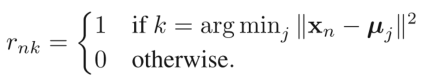
隐变量(latent variable):无法被直接观测到，需要通过模型和观测变量进行推断。利用潜变量来解释观测变量的数学模型，称为潜变量模型，GMM和HMM都是潜变量模型。潜变量模型将不完全数据(只有观测数据)的边缘分布转换成容易处理的完全数据(观测数据+潜变量)的联合分布。



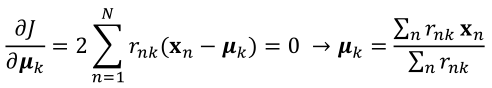
在上面这个例子里，潜变量是类别，是未知的，观测变量是数据点。



第一步，选择初始的uk，并保持uk固定，最小化J关于rnk (E步)

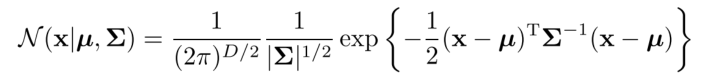


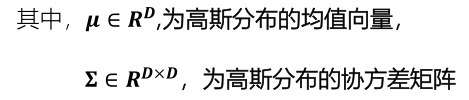
第二步，保持rnk固定，最小化J关于uk (M步，重新估算参数)



GMM模型

D维随机变量的高斯分布为：

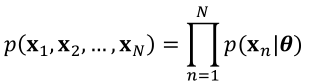




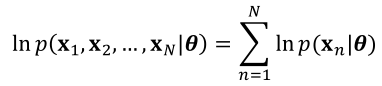
选择高斯分布的原因：高斯分布在自然界的数据广泛存在。在适当的条件下，大量相互独立随机变量的均值经适当标准化依分布收敛于正态分布。

最大似然估计：

假设随机变量X服从分布，其中，为待估计的参数，如果可以获得N个互相独立的X的采样点{}，则似然函数的定义为：



在实际使用中，一般采用对数似然函数：

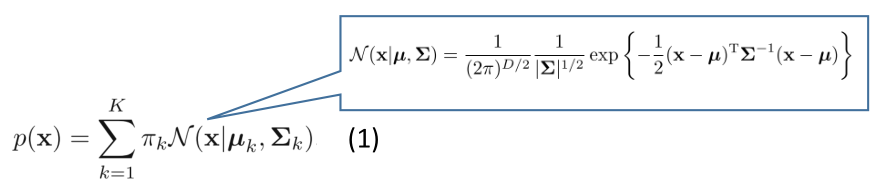


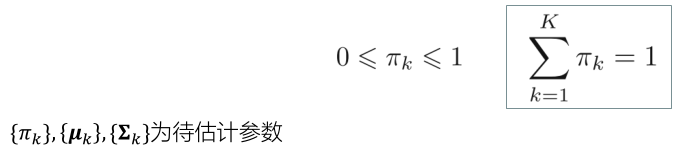
参数的最大似然估计为：

IMG_256

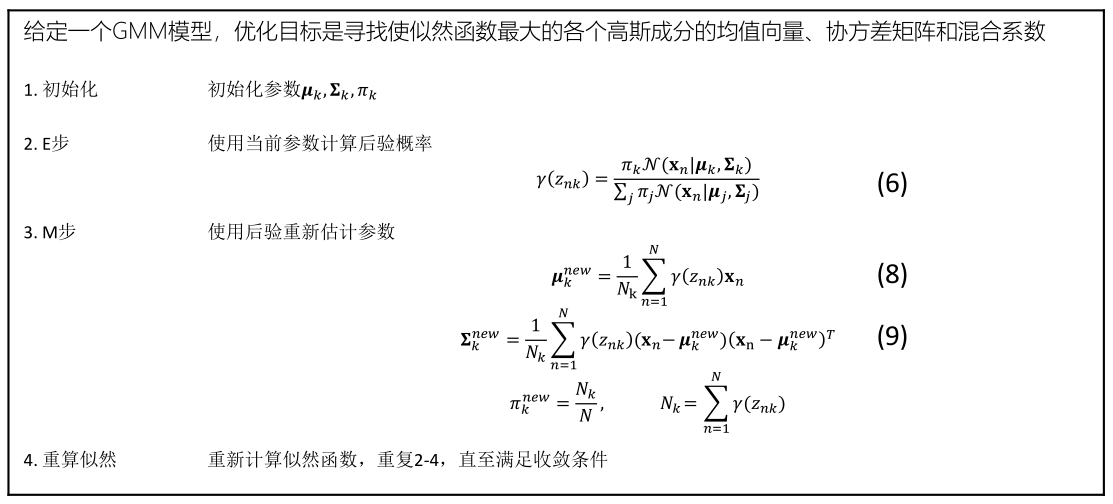
最大似然估计的一个直观解释就是：现在有一个袋子，里面放有红、白两种颜色的球共100个，其中一种颜色的球有99个，另一种颜色的球1个。从袋子中任取一球，结果为红球，我们有理由认为袋子里有99个红球。

高斯混合分布：直观的说就是几个单高斯以一定权重混合在一起

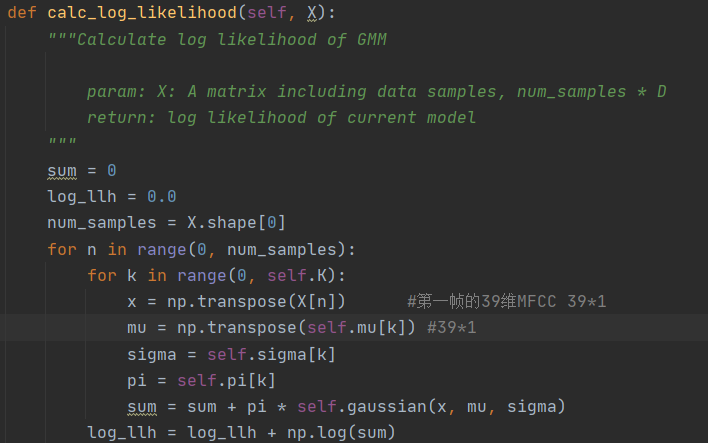




GMM模型参数估计的EM算法：



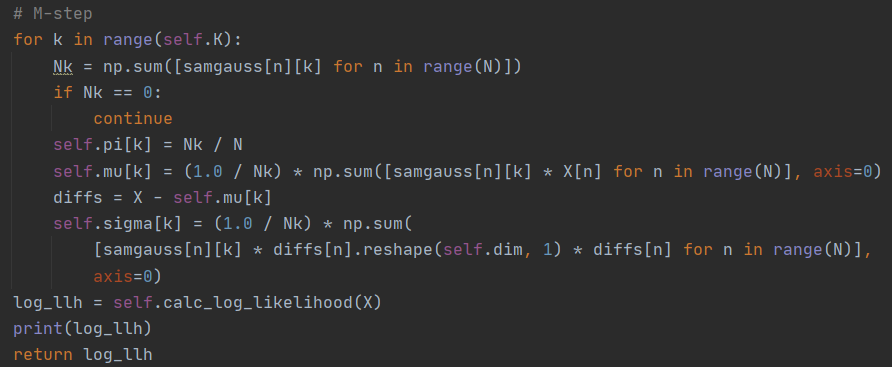
Step1：似然概率计算函数

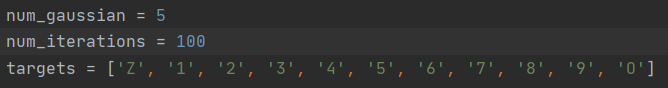


Step2：后验重新估计参数函数

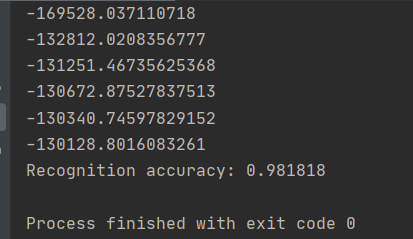


Step3：M步，重新更新三个参数





迭代100次，识别‘Z’‘1’到‘9’



孤立词识别正确率达到98.18%