

声纹识别

概念

说话人识别，判断一段声音是谁说的。

特征

一般使用 plp 或者 mfcc 做帧的特征抽取，其中帧往往是采样获得的，不是每个都取。

LBG 建模

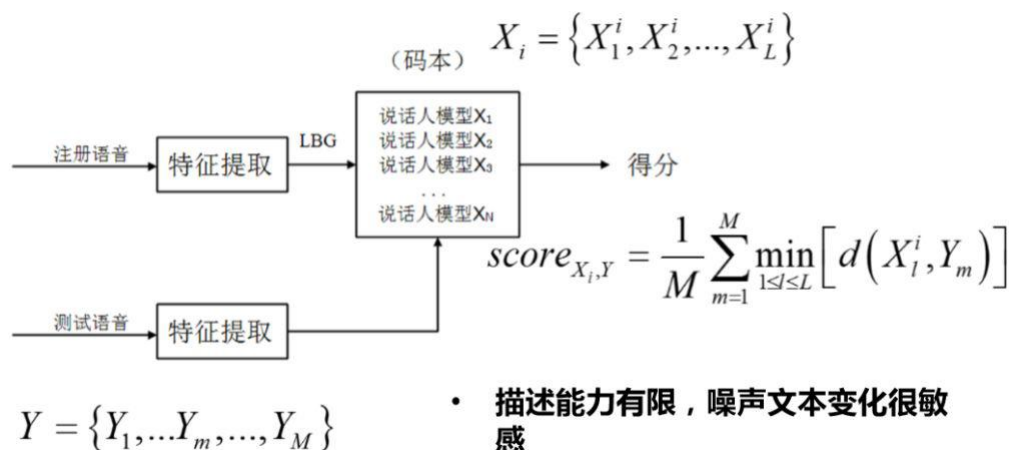
根据对机器学习的常识，首先就会想到一种方法来做声纹识别：

假如已经训练好了说话人模型：一个人对应一个模型。

伪代码：

- 1) 对帧进行采样
- 2) 对帧进行特征抽取，记为 y_1, y_2, \dots, y_n
- 3) 将所有对特征向量代入每个说话人模型，计算所有对 y_i 和说话人模型之间对距离之和，距离最小对 score 对应对说话人模型就是这段语音对应对模型。

– LBG (Linde–Buzo–Gray Algorithm)



缺点：

音频文件避免不了各种背景噪音。

就是音频里面往往又很多噪音，这些噪音会影响距离对大小。

高斯混合模型

与语音识别中对 GMM 用法不一样，声纹识别中的 GMM 是对一段语音中的帧会抽取特征，

如 mfcc，一段语音会有多帧，将这些帧的语音特征放在一起训练一个 GMM。而声学模型是对音素的某一个状态训练 gmm 模型。

如果使用 mfcc 抽取特征，则为 13 维的特征向量。

这个时候高斯分布的维度也是 13 维，而高斯的数量是一个超参数可以调整。

训练高斯混合模型的时候，首先参数需要一个初始值，然后使用 em 算法逐渐收敛。

初始值的选取方法是：

使用 kmeans 对数据进行聚类，假如高斯分布对数量设为 m，则使用 kmeans 聚 m 类。然后对每个类求解高斯模型参数。这就是高斯混合模型对初始值。

GMM-UBM

思想：

其实 UBM 就是 GMM 模型，只是训练的目的不同，GMM 我们希望训练得到一个能够表征说话人音素分布的模型，而 UBM 是希望得到一个通用的模型，简单的说就是能够反应所有人共性的模型，其实某种意义上说就是一个取均值的过程。

操作方法：

对所有人对应的音频混杂在一起训练一个高斯混合模型。这个时候训练出来的高斯混合模型我们理解为“通用模型”

在通用模型上面进行微调，就可以得到每个人的模型。

MAP 自适应过程

虽然高斯混合模型的参数为四个：

$$\theta = (w_i, u_i, \sigma_i), i = 1, 2, \dots, C$$

和协方差矩阵。但是协方差矩阵一般设置为对角阵。

C 为 GMM 的混合阶数；说话人 X 的训练语音的特征向量序列为 x

1. 首先计算语音特征向量序列中的各个向量相对于每个 UBM 混元的概率得分。

$$p(O | \phi) = \frac{1}{(2\pi)^{d/2} \sigma^2} \exp \left[-\frac{1}{2} \sum_{t=1}^T \left(\frac{O_t - \phi}{\sigma} \right)^2 \right]$$

2. 对于 UBM 中的任意混元 i，特征向量 x_i 对于它的后验分布概率为：

$$p(i | x_i, \lambda_\Omega) = \frac{\omega_i p(x_i | \mu_i, \sum_i)}{\sum_{j=1}^C \omega_j p(x_i | \mu_j, \sum_j)}$$

3. 利用后验概率计算均值所需要的统计量

$$p(i | \lambda_{\Omega}) = \sum_{t=1}^T p(i | x_t, \lambda_{\Omega})$$
$$E_i(X) = \frac{1}{p(i | \lambda_{\Omega})} \sum_{t=1}^T p(i | x_t, \lambda_{\Omega}) x_t$$

4. 最后利用上面两个统计量对 UBM 均值进行更新，其对任意混元 i 的均值更新表达式如下：

$$\hat{\mu}_i = \partial_i E_i(X) + (1 - \partial_i) \mu_i$$

自适应 ∂ 系数控制着旧估计与新估计之间的均衡，自适应算法就是对 UBM 参数做个微调，使得参数在一定背景的基础下调整到能够表征说话人发音特征，在语音数据不充分的情况下，没有覆盖到的发音特征可以用 UBM 的平均发音特征来代替。第 2 步公式，反应了当前模型下，第 j 个观测数据，来自第 K 个分模型的概率，称为分模型 K 对观测数据 y_j 的响应度。

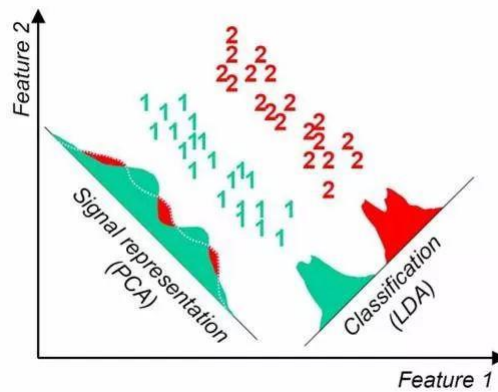
总结:

使用所有特征训练一个高斯混合模型（通用模型），使用 MAP 获得每一段语音对应的高斯模型的均值参数。这个均值向量就可以表示这段语音的声纹特征。

LDA(线性判别分析)

LDA 的思想

LDA 是一种监督学习的降维技术，也就是说它的数据集的每个样本是由类别输出的。PCA 是不考虑样本类别输出的无监督降维技术。LDA 的思想可以用一句话概括，就是“投影后类内方差最小，类间方差最大”，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据中新之间的距离尽可能的大。



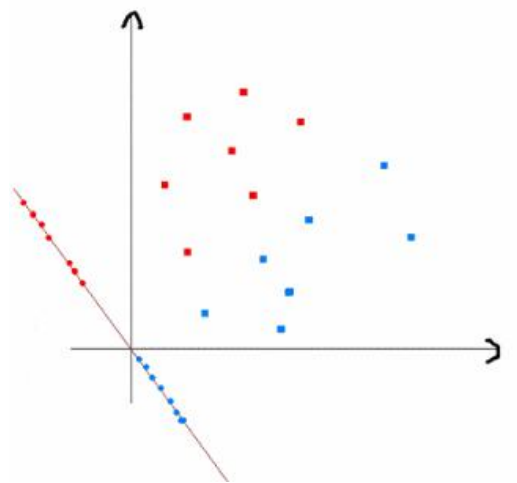
LDA 的全称是 Linear Discriminant Analysis (线性判别分析)，**是一种 supervised learning**。有些资料上也称为是 Fisher's Linear Discriminant，因为它被 Ronald Fisher 发明自 1936 年，Discriminant 这个词我个人的理解是，一个模型，不需要去通过概率的方法来训练、预测数据，比如说各种贝叶斯方法，就需要获取数据的先验、后验概率等等。

LDA 的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。要说明白 LDA，首先得弄明白线性分类器(Linear Classifier)：因为 LDA 是一种线性分类器。对于 K-分类的一个分类问题，会有 K 个线性函数：

$$y_k(x) = w_k^T x + w_{k0}$$

当满足条件：对于所有的 j，都有 $Y_k > Y_j$ 的时候，我们就说 x 属于类别 k。对于每一个分类，都有一个公式去算一个分值，在所有的公式得到的分值中，找一个最大的，就是所属的分类了。

上式实际上就是一种投影，是将一个高维的点投影到一条高维的直线上，LDA 最求的目标是，给出一个标注了类别的数据集，投影到了一条直线之后，能够使得点尽量按类别区分开，当 k=2 即二分类问题的时候，如下图所示：



红色的方形的点为 0 类的原始点、蓝色的方形点为 1 类的原始点，经过原点的那条线就是投影的直线，从图上可以清楚的看到，红色的点和蓝色的点被**原点**明显的分开了，这个数据只是随便画的，如果在高维的情况下，看起来会更好一点。下面我来推导一下二分类 LDA 问题的公式：

假设用来区分二分类的直线（投影函数）为：

$$y = w^T x$$

LDA 分类的一个目标是使得不同类别之间的距离越远越好，同一类别之中的距离越近越好，所以我们需要定义几个关键的值。

类别 i 的原始中心点为：（ D_i 表示属于类别 i 的点）
$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

类别 i 投影后的中心点为：

$$\widetilde{m}_i = w^T m_i$$

衡量类别 i 投影后，类别点之间的分散程度（方差）为：

$$\widetilde{s}_i = \sum_{y \in Y_i} (y - \widetilde{m}_i)^2$$

最终我们可以得到一个下面的公式，表示 LDA 投影到 w 后的损失函数：

$$J(w) = \frac{|\widetilde{m}_1 - \widetilde{m}_2|^2}{\widetilde{s}_1 + \widetilde{s}_2}$$

我们分类的目标是，使得类别内的点距离越近越好（集中），类别间的点越远越好。分母表示每一个类别内的方差之和，方差越大表示一个类别内的点越分散，分子为两个类别各自的中心点的距离的平方，我们最大化 $J(w)$ 就可以求出最优的 w 了。想要求出最优的 w，可以使用拉格朗日乘子法，但是现在我们得到的 $J(w)$ 里面，w 是不能被单独提出来的，我们就得想办法将 w 单独提出来。

我们定义一个投影前的各类别分散程度的矩阵，这个矩阵看起来有一点麻烦，其实意思是，如果某一个分类的输入点集 D_i 里面的点距离这个分类的中心点 m_i 越近，则 S_i 里面元素的值就越小，如果分类的点都紧紧地围绕着 m_i ，则 S_i 里面的元素值越更接近 0。

$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^T$$

带入 S_i ，将 $J(w)$ 分母化为：

$$\tilde{s}_i = \sum_{x \in D_i} (w^T x - w^T m_i)^2 = \sum_{x \in D_i} w^T (x - m_i)(x - m_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T (S_1 + S_2) w = w^T S_w w$$

同样的将 $J(w)$ 分子化为：

$$|\tilde{m}_1 - \tilde{m}_2|^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w$$

这样损失函数可以化成下面的形式：

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

这样就可以用最喜欢的拉格朗日乘子法了，但是还有一个问题，如果分子、分母是都可以取任意值的，那就会使得有无穷解，我们将分母限制为长度为 1（这是用拉格朗日乘子法一个很重要的技巧，在下面将说的 PCA 里面也会用到，如果忘记了，请复习一下高数），并作为拉格朗日乘子法的限制条件，带入得到：

$$\begin{aligned} c(w) &= w^T S_B w - \lambda(w^T S_w w - 1) \\ \Rightarrow \frac{dc}{dw} &= 2S_B w - 2\lambda S_w w = 0 \\ \Rightarrow S_B w &= \lambda S_w w \end{aligned}$$

这样的式子就是一个求特征值的问题了。

对于 $N(N>2)$ 分类的问题，我就直接写出下面的结论了：

$$S_W = \sum_{i=1}^c S_i$$

$$S_B = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T$$

$$S_B w_i = \lambda S_W w_i$$

这同样是一个求特征值的问题，我们求出的第 i 大的特征向量，就是对应的 W_i 了。