ASR定义

定义：

- **什么是语音识别？**
  输入语音信号其对应的输出文本序列。
- **如何形式化这个问题？**
  现假设长度为D的语音/特征序列为$S_D = (s_1, s_2, s_3, \dots s_d)$，单词/音素序列集合为W。需要知道语音对应的最佳单词/音素序列$W_L^*(w_1, w_2, w_3, \dots w_l)$。据此可以建立概率模型：

$$W^* = argmax\{P(W|S)\}$$

根据贝叶斯公式：

$$W^* = argmax\frac{P(S|W)P(W)}{P(S)}$$

因为S已知所以$P(S)$为常量：

$$W^* = argmax P(S|W)P(W)$$

建模：

- **目标是什么？**
  需要得到条件概率概率(似然概率)$P(S|W)$和先验概率$P(W)$的最大乘积。
  - 对于$P(W)$来说单词\因素序列的条件概率可以通过语言模型的统计可得。
  - $P(S|W)$的含义是：给定单词序列W，得到特定音频信号S的概率。
- **如何建模？**
  - 建模单元
    由于单词数太多用来建模的话参数过多，W序列因此通常使用更小的建模单元比如：字词、音节、或者音素。

由于语音的短时不变特性，所以通常从一个短时窗口(20ms~50ms)提取一次特征(MFCC, FBANK, PLP, PITCH等)。因此S序列通常使用20ms内特征作为建模单元。

- 建模工具
  生成模型是对类条件密度建模，可以直接应用在该问题上。
  判别模型则直接对$P(W|S)$建模需要除以状态的先验概率得到似然度。
  生成模型和判别模型的对比

GMM

原理：

- **定义**
  高斯混合模型是单一高斯概率密度函数的延伸，由多个高斯概率密度函数组合而成，是将变量分布分解为若干基于高斯概率密度函数分布的统计模型。由于混合高斯分布的多模态M，因此其足以描述显示出多模态性质的物理数据(比如语音特征的主要分量)。其原理类似于三角函数的组合拟合任意曲线。
  假定其符合正太分布，μ为均值，σ^2为方差。
  其一维密度函数为：

$$p(x)=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]=N(x;\mu,\sigma^2)$$
$$(-\infty,+\infty;\sigma^2)$$

高维(D维)密度函数为：

$$p(x)=\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}exp[-\frac{1}{2}(x-\mu)^T\Sigma^-1(x-\mu)]=N(x;\mu,\Sigma)$$

混合高斯的概率密度函数为：

$$p(x)=\sum_{m}^{M}\frac{c_m}{(2\pi)^{\frac{1}{2}}\sigma_m}exp[-\frac{1}{2}(\frac{x-\mu_m}{\sigma_m})^2]=\sum_{m=1}^{M}N(x;\mu_m,\sigma_m^2)$$

其中混合权重为正实数，和为1：$\sum_{m=1}^{M}c_m=1$
将其扩展到多变量的多元混合高斯分布，其联合概率密度为：

$$p(x)=\sum_{m=1}^{M}\frac{c_m}{(2\pi)^{\frac{D}{2}}|\Sigma_m|^{\frac{1}{2}}}exp[-\frac{1}{2}(x-\mu_m)^T\Sigma_m^{-1}(x-\mu_m)]=\Sigma_{m-1}^{M}c_mN(x;\mu_m,\Sigma_m)$$

建模：

- 目标是什么？
  根据上述定义，目标已经很明确了，就是估计混合高斯分布包含的一些参数变量$\Theta=\{c_m,\mu_m,\Sigma_m\}$。
  这种通过训练样本估计概率密度，再通用统计决策进行类别鉴定的方法称为基于样本的两步贝叶斯决策。模型已确定，根据样本分布估计参数问题，一般采用极大似然估计或者贝叶斯估计。两者区别有些哲学含义，类似可知论和不可知论。但是考虑到计算量此处用极大似然估计。

$$l(\theta)=P(X|\theta)=P(x_1,x_2,...,x_N)|\theta)=\Pi_{i=1}^{N}P(x_i|\theta)$$

$$\begin{aligned}\theta_{MLP}&=argmax\ l(\theta)\\&=argmax\ log\,l(\theta)\\&=argmax\ \sum logP(x_i|\theta)\end{aligned}$$

- 如何训练？：
  多数情况下极大似然参数θ没有显式解，因此EM算法就出现了，EM算法是解决极大似然估计的迭代方法。
  X是输入参数，Z是隐含变量，θ是模型参数。
  由于还有隐藏的观测值数据，根据边缘分布的定义：

$$\theta=argmax\ \sum logP(x_i|\theta)=argmax\ \sum log\sum P(x_i,z_i|\theta)$$

EM公式:

$$\theta^{(t+1)} = \underset{\theta}{argmax} \int_Z log\, P(X, Z|\theta).P(Z|X, \theta^{(t)})dZ$$

E-step:
$log\, p(X, Z|\theta)$乘$p(Z|X, \theta^{(t)})$再对Z积分，即$E_{Z|X,\theta^{(t)}}[log\, p(X, Z|\theta)]$.
M-step:

$$\theta^{(t+1)} = \underset{\theta}{argmax}\, E_{Z|X,\theta^{(t)}}[log\, p(X, Z|\theta)]$$

证明过程可见GMM的推导及实现。

- 怎么使用？：
  模型收敛之后得到的模型，通过比较不同类别结果的大小选择概率最大的类别。到这里配合上帧级别的音素标签的话已经算是一个语音识别系统了。但是语音是一种前后关联性很强的序列，完全忽略掉时序信息不仅无法做到准确而且会丢失有用的信息。

讨论：

- GMM发展：
  技术从来都不是一成不变的，自1995年GMM被用于语音识别之后就一直有发展，其中包括训练中收敛速度的优化，而且近年来随着无监督学习的发展使得GMM再次被关注。
- GMM有什么缺点？
  - 首先，高维度情况下表现不好，特别是样本量不足时协方差的估计会很困难；
  - 其次，GMM是一系列高斯分布的组合，但应使用多少分布是未知的，需要用户自行定义或调试；
  - GMM没考虑语音的序列信息。
  - GMM复杂度较高，对呈非线性或近似非线性地数据而言合适地模型仅需少量地参数，但是GMM需要的参数很多。

GMM-HMM

原理：

- 定义：
  HMM(Hidden Markov Model), 也称隐性马尔科夫模型，是一个概率模型，用来描述一个系统隐性状态的转移和隐性状态的表现概率，属于生成模型。针对语音特征的长度可变性，HMM能够很好地生成序列模型。
- 形式化：
  如图所示，HMM包含两个部分状态序列和观察序列。
  隐马尔科夫模型由初始状态概率向量π、状态转移概率矩阵A 以及观测概率矩阵B决定。因此隐马尔科夫模型λ可以用三元符号表示，即 ：$\lambda = (A, B, \pi)$ 。其中A, B, π称为隐马尔科夫模型的三要素：
    - 初始状态概率向量$\pi = [\pi_i]$, $i = 1, 2, ..., N$,其中，$\pi_i = P(q_1 = i)$，代表t=1时状态q等于i的概率。
    - 状态转移概率矩阵$A = [a_{ij}$, $i, j = 1, 2, ..., N]$，$a_{ij} = P(q_t = j|q_{t-1} = i)$，$i, j = 1, 2, ..., N$，代表t时刻状态$q$为$i$情况下，转移到t+1时刻状态$q$为$j$的概率。
    - 观测概率矩阵$B = [b_i(k)]$，$b_i(k) = P(o_t = v_k|q_t = i)$，代表t时刻状态$q$为$i$情况下，观测$o$为$v_k$的概率。如果概率密度是连续的，比如GMM-HMM，则$b_i(o_t) = \sum_{m=1}^{M} \frac{c_{i,m}}{(2\pi)^{D/2}|\Sigma_{i,m}|^{1/2}} exp[-\frac{1}{2}(o_t - \mu_{i,m})^T \Sigma_{i,m}^{-1}(o_t - \mu_{i,m})]$。
    - 状态转移概率矩阵和初始状态概率向量 确定了隐藏的马尔科夫链，生成不可观测的状态序列。
    - 观测概率矩阵 确定了如何从状态生成观测，与状态序列一起确定了如何产生观测序列。

  从定义可知，隐马尔科夫模型做了两个基本假设：
    - 齐次性假设：即假设隐藏的马尔科夫链在任意时刻 的状态只依赖于它在前一时刻的状态，与其他时刻的状态和观测无关，也与时刻 无关，即：

      $$P(i_t|i_{t-1}, o_{t-1}, ..., i_1, o_1) = P(i_t|i_{t-1}),\ t = 1, 2, ..., T$$

    - 观测独立性假设，即假设任意时刻的观测值只依赖于该时刻的马尔科夫链的状态，与其他观测及状态无关，即：

$$P(O_t|i_t,o_t,...,i_{t+1},o_{t+1}.i_t,i_{t-1}o_{t-1},...,i_1,o_1)=P(o_t|i_t),\ t=1,2,...,T$$

- 基本问题推导过程：
    - 概率计算问题：给定模型$\lambda=(A,B,\pi)$和观测序列$O=o_1,o_2,...,o_T$，计算观测序列O出现的概率$P(O:\lambda)$。即：评估模型λ与观察序列O之间的匹配程度。可使用(Forward-Backword)前向后向算法。
    - 学习问题：已知观测序列$O=o_1,o_2,...,o_T$，估计模型$\lambda=(A,B,\pi)$的参数，使得在该模型下观测序列概率$P(O:\lambda)$最大。即：用极大似然估计的方法估计参数。可使用(Viterbi)维特比算法。
    - 预测问题（也称为解码问题）：已知模型$\lambda=(A,B,\pi)$和观测序列$O=o_1,o_2,...,o_T$，求对给定观测序列的条件概率$P(I|O)$最大的状态序列$I=(i_1,i_2,...,i_T)$。即：给定观测序列，求最可能的对应的状态序列。可使用(Baum-Welch)鲍姆-韦尔奇算法
- 在语音识别任务中，观测值为语音信号，隐藏状态为文字。解码问题的目标就是：根据观测的语音\特征信号S_D来推断最有可能的文字序列W_L。

建模：

- 目标是什么？
  HMM作为生成模型，参考上文GMM(GMM可视为HMM的特殊情况)，可以使用极大似然估计参数$\lambda=(A,B,\pi)$。HMM中连续观测向量的概率分布由概率密度函数描述(PDF)，GMM-HMM中概率密度函数为多元混合高斯分布。
  此处类似前文，将参数包含在θ内，极大似然估计函数为$argmax\ \sum logP(x_i|\theta)$。同样模型存在隐藏变量$Z$，与GMM不同的是此处的$Z$符合马尔科夫链，因此观测变量和隐藏变量的联合概率分布函数为：
  $$P(O,Q|\theta)=P(q_1|\pi)[\Pi_{n=2}^{N}P(q_n|q_{n-1},A)]\Pi_{m=1}^{N}P(o_m|q_{mv},B)$$
- 如何训练？
  上文提到EM算法是种通用解决最大化似然度估计的方法，所以HMM也使用EM算法，但是当隐藏变量符合马尔科夫链的形式时，EM算法可以推到为Baum-Welch算法(EM算法的特例)。
  E-step:
  M-step:
- 怎么使用？

讨论：

DNN-HMM

原理：

建模：

讨论：

DNN-CTC

原理：

建模：

讨论：

TRANSFORMER

原理：

建模：

讨论：

讨论