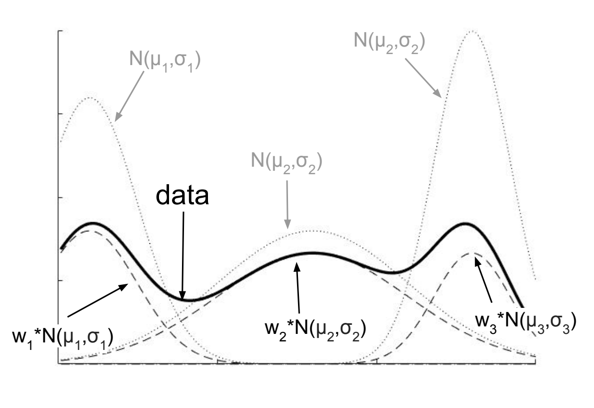
K-means, Gaussian Mixture Models, UBM-GMM

Pattern Recognition, Jul-Nov 2019

Indian Institute of Technology Madras

September 6, 2019

Gaussian mixture models (GMM)



- A GMM is the weighted sum of individual Gaussian distributions

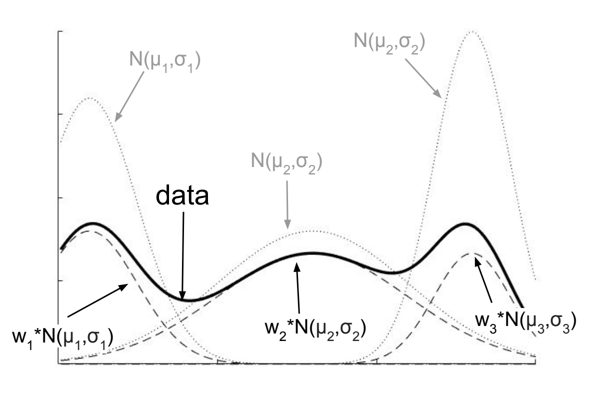
$$P(\vec{x}|\lambda_s) = \sum_{k=1}^{M} w_k \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)$$

$$\lambda_s = \{w_k, \vec{\mu}_k, \Sigma_k\}_{k=1}^{M}$$

$$\sum_{k=1}^{K} w_k = 1$$

$$\mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) = (2\pi)^{-d/2} \left|\Sigma_k\right|^{-1/2} exp\, \{-\frac{1}{2}(\vec{x} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{x} - \vec{\mu}_k)\}$$

Gaussian mixture models (GMM)



- The problem of fitting a GMM is a incomplete data problem. Hence, the mixtures needs to be estimated iteratively using Expectation Maximization (E-M) algorithm.

Parameter estimation of GMM using E-M algorithm

- Estimate parameters $(\bar{\mu}_k, \Sigma_k)$.
- Define one more quantity 'responsibility'.

$$P(z_k = 1|x) = \frac{P(z_k = 1)P(\bar{x}|z_k = 1)}{\sum_{j=1}^{K} P(\bar{x}|z_j)P(z_j = 1)} \tag{1}$$

$P(z_k = 1|x) = \gamma_k$, which is the *responsibility* of k_{th} mixture in describing a point x.

$$P(z_k = 1|x) = \frac{P(z_k = 1)P(\bar{x}|z_k = 1)}{\sum_{j=1}^{K} P(\bar{x}|z_j)P(z_j = 1)}$$

$$= \frac{\pi_k N(\bar{x}; \bar{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\bar{x}; \bar{\mu}_j, \Sigma_j)}$$

$$N(\bar{x}, \bar{\mu}_j, \Sigma_j) = \frac{-1}{(\sqrt[2]{2\pi})^d |\Sigma|} e^{\left(\frac{-1}{2}(\bar{x} - \bar{\mu}_j)^t \Sigma^{-1} (\bar{x} - \bar{\mu}_j)\right)}$$

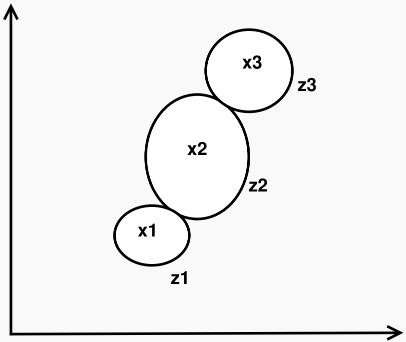
A single point will be described completely by all the mixtures together.

$$\sum_{k=1}^{K} \gamma_k = 1$$

$$\gamma_{nk} = P(z_k = 1|\bar{x}_n)$$

$$\sum_{n=1}^{N} \gamma_{nk} = N_k$$

N_k is the effective number of points that belongs to k^{th} cluster



$$\gamma(z1=1|\bar{x}2)=.125,\quad \gamma(z1=1|\bar{x}1)=.75,\quad \gamma(z1=1|\bar{x}3)=.05$$
$$\gamma(z2=1|\bar{x}2)=.75,\quad \gamma(z1=1|\bar{x}1)=.20,\quad \gamma(z2=1|\bar{x}3)=.20$$
$$\gamma(z3=1|\bar{x}2)=.125,\quad \gamma(z1=1|\bar{x}1)=.05,\quad \gamma(z3=1|\bar{x}3)=.75$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \gamma_{nk} \bar{x}_n$$

Estimation of parameters

$$\theta_{ML} = \arg\max_{\theta} \log p(D|\theta)$$

$$\bar{\theta} = [\bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_k, \Sigma_1, \Sigma_2, ...\Sigma_k, \pi_1, \pi_2, ..., \pi_k]$$

$$\ln P(D|\theta) = \ln \Pi_{n=1}^{N} P(\bar{x}_n|\theta) = l(\theta)$$

$$lnP(D|\theta) = \sum_{n=1}^{N} \ln \sum_{j=1}^{K} \pi_j N(\bar{x}_n; \bar{\mu}_j, \Sigma_j)$$

$$\frac{\partial l(\bar{\theta})}{\partial \bar{\theta}} = 0$$

$$\frac{\partial l(\bar{\theta})}{\partial \bar{\mu}_k} = 0; \frac{\partial l(\bar{\theta})}{\partial \bar{\Sigma}_k} = 0$$

π_k requires a constraint $\sum_{k=1}^{K} \pi_k = 1$

$$\frac{\partial l(\bar{\theta})}{\partial \bar{\mu}_k} = \sum_{n=1}^{N} \frac{\partial}{\partial \bar{\mu}_k} \ln \sum_{j=1}^{K} \pi_j N(\bar{x}_n; \bar{\mu}_j, \Sigma_j)$$

$$= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{j=1}^{K} \pi_j N(\bar{x}_n; \bar{\mu}_j, \Sigma_j)} \frac{\partial}{\partial \bar{\mu}_k} N(\bar{x}_n; \bar{\mu}_k, \Sigma_k)$$

$$\frac{\partial}{\partial \bar{\mu}_k} N(\bar{x}_n; \bar{\mu}_k, \Sigma_k) = \frac{-1}{(\sqrt[2]{2\pi})^d |\Sigma|^{\frac{1}{2}}} e^{(\frac{-1}{2}(\bar{x}_n - \bar{\mu}_k)^t \Sigma_k^{-1} (\bar{x}_n - \bar{\mu}_k))}$$

$$\frac{\partial}{\partial \bar{x}} \bar{x}^t M \bar{x} = 2M\bar{x}$$

$$\frac{\partial l(\bar{\theta})}{\partial \bar{\mu}_k} = -N(\bar{x}_n; \bar{\mu}_k, \Sigma_k)\Sigma_k^{-1}(\bar{x}_n - \bar{\mu}_k)$$

$$= \textstyle\sum_{n=1}^{N} \frac{\pi_k N(\bar{x}; \bar{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\bar{x}; \bar{\mu}_j, \Sigma_j)}$$

$$\frac{\partial}{\partial \bar{\mu}_k} = 0 \implies \sum_{n=1}^{N} \gamma_{nk} \Sigma_k^{-1}(\bar{x}_n - \bar{\mu}_k) = 0$$

$$\sum_{n=1}^{N} \gamma_{nk} \bar{x}_n = \sum_{n=1}^{N} \gamma_{nk} \bar{\mu}_k$$

$$\bar{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \bar{x}_n$$

To find Σ_k, take derivative with respect to Σ_k and equate it to 0.

$$\bar{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\bar{x}_n - \hat{\mu}_k)(\bar{x}_n - \hat{\mu}_k)^t$$

Identities

- $\frac{\partial |A|}{\partial A} = |A| A^{-1}$
- $\frac{\partial \bar{\mu}^t A^{-1} \bar{\mu}}{\partial A} = A^{-1} \bar{\mu} \bar{\mu}^t A^{-1}$
- $\frac{\partial (\bar{u}.\bar{v})}{\partial x} = \bar{v}^t \frac{d\bar{u}}{d\bar{x}} + \bar{u}^t \frac{d\bar{v}}{d\bar{x}}$

Estimation of π_k such that $\sum_{j=1}^{K} \pi_j = 1$

This is a constraint optimization problem.

$$L(\pi_k, \lambda) = \sum_{n=1}^{K} \ln \sum_{j=1}^{K} \pi_j N(\bar{x}_n; \bar{\mu}_j, \Sigma_j, \pi_j) - \lambda(\sum_{j=1}^{K} \pi_j - 1)$$

$$\frac{\partial L(\pi_k, \lambda)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\pi_k N(\bar{x}; \bar{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\bar{x}; \bar{\mu}_j, \Sigma_j)}$$

$$\lambda \pi_k = \sum_{n=1}^{N} \gamma_{nk}$$

$$\lambda \sum_{k=1}^{K} \pi_k = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}$$

$$\pi_k = \frac{N_k}{N}$$

GMM : EM algorithm (Recap)

1) Random initialization of $\vec{\mu}_k, \Sigma_k$ and w_k

2) Expectation-Step

Align vectors to model

$$\gamma_{nk} = \frac{w_k \mathcal{N}(\vec{x}_n | \vec{\mu}_k, \Sigma_k)}{\sum_{j=1}^{M} w_j \mathcal{N}\left(\vec{x}_n | \vec{\mu}_j, \Sigma_j\right)}$$

$$N_k = \sum_{n=1}^{N} \gamma_{nk}$$

$\rightarrow \gamma_{nk}$ is the responsibility of k-th component towards n-th feature vector

3) Maximization-Step

Update model parameters by maximum likelihood estimation (MLE)

$$\hat{\vec{\mu}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \vec{x}_n$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\vec{x}_n - \hat{\vec{\mu}}_k)(\vec{x}_n - \hat{\vec{\mu}}_k)^T$$

$$\hat{w}_k = \frac{N_k}{N}$$

4) Repeat step 2 and 3 until convergence.

GMM-proof of convergence

- Let $\bar{\Theta}^{\text{old}}$ be parameters used at the start of any EM-iteration and $\bar{\Theta}$ be the updated parameters.
- Let $\mathcal{D}$ be the given set of data points and $\bar{z}$ be the latent variable defined by $\bar{\Theta}^{\text{old}}$.
- We need to prove that $P\left(\mathcal{D} \mid \bar{\Theta}\right) \geq P\left(\mathcal{D} \mid \bar{\Theta}^{\text{old}}\right)$ for every iteration.
- In every iteration we need to maximize $E\left[log\ P\left(\mathcal{D}, \bar{z} \mid \bar{\Theta}\right) \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right]$
- Lets define the auxiliary function as follows:

$$
\begin{aligned}
A\left(\bar{\Theta}, \bar{\Theta}^{\text{old}}\right) &= \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)\ log\ P\left(\mathcal{D}, \bar{z} \mid \bar{\Theta}\right) \\
&= \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)\ log\ \left(P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}\right) P\left(\mathcal{D} \mid \bar{\Theta}\right)\right) \\
&= \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)\ log\ \left(P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}\right)\right) \\
&+ \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)\ log\ \left(P\left(\mathcal{D} \mid \bar{\Theta}\right)\right)
\end{aligned}
$$

GMM-proof of convergence (Contd..)

- Substituting $\bar{\Theta}^{\text{old}}$ for $\bar{\Theta}$ in the previous equation we get

$$A\left(\bar{\Theta}^{\text{old}}, \bar{\Theta}^{\text{old}}\right) = \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right) \; log \; \left(P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)\right) + \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right) \; log \; \left(P\left(\mathcal{D} \mid \bar{\Theta}^{\text{old}}\right)\right)$$

- Now $log \; P\left(\mathcal{D} \mid \bar{\Theta}\right) - log \; P\left(\mathcal{D} \mid \bar{\Theta}^{\text{old}}\right)$ can be written as:

$$log \; P\left(\mathcal{D} \mid \bar{\Theta}\right) - log \; P\left(\mathcal{D} \mid \bar{\Theta}^{\text{old}}\right) = A\left(\bar{\Theta}, \bar{\Theta}^{\text{old}}\right) - A\left(\bar{\Theta}^{\text{old}}, \bar{\Theta}^{\text{old}}\right) + \sum_{z} P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right) log \left(\frac{P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}^{\text{old}}\right)}{P\left(\bar{z} \mid \mathcal{D}, \bar{\Theta}\right)}\right) \tag{2}$$

GMM-proof of convergence (Contd..)

In Equation 2

- The last part of the equation is *Kullback-Leibler* divergence which is always positive or null.
- The update equation are derived such that the A increases (See https://courses.iitm.ac.in/mod/resource/view.php?id=789 for derivation based on auxiliary function)

Therefore, the likelihood always increases from iteration to iteration.

UBM-GMM

Universal background model (UBM)

- UBM is a GMM trained on huge data pooled together from all the available classes.
- To overcome the huge data requirement for training GMM for individual classes.
- UBM is a GMM built with all class data using MLE algorithm.
- Maximum A-Posteriori (MAP) adaptation is used to train target models.
- UBM does not discriminate different speaker but acts as a reference for all class models.

UBM-GMM : MAP adaptation (contd...)

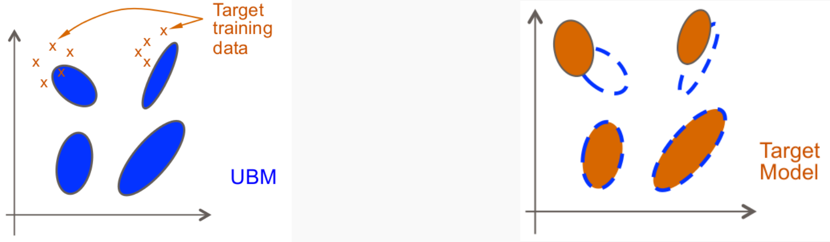


Figure 1: Adaptation using MAP

- Target model mean is updated using sufficient statistics, mixing co-efficient (α_k) and relevance factor r as

$$\alpha_k = \frac{N_k}{N_k + r}$$

$$\hat{\vec{\mu}}_k^{new} = \alpha_k \hat{\vec{\mu}}_k + (1 - \alpha_k)\vec{\mu}_k^{ubm}$$

UBM-GMM : MAP adaptation (contd...)

Figure 2: Adaptation using MAP

- The covariance can also be adapted using the same formule. But for most of the application we limit ourselves to adapting the mean.

Scoring in UBM-GMM

UBM acts as a reference for all class models. Hence the scores are calcluated with refence to the UBM.

Likelihood scoring (in the context of speech)

The average log-likelihood ratio score for a test utterance $\mathcal{X} = \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_T\}$ with claim is calculated as

$$LogLikelihood_{avg}(\mathcal{X}, \lambda_{claim}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{log\ P(\bar{\mathbf{x}}_t|\lambda_{claim}) - log\ P(\bar{\mathbf{x}}_t|\lambda_{UBM})\}$$

Advantages of MAP in UBM-GMM

- Mixtures of UBM and target GMM have one-one correspondence
- Only few mixture components contribute to the a paticular class's feature vectors
- $LogLikelihood_{avg}$ can be computed with maximum contributing C mixture components of the UBM.