Statistical Machine Learning from Data

Gaussian Mixture Models

Samy Bengio

IDIAP Research Institute, Martigny, Switzerland, and
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
bengio@idiap.ch
http://www.idiap.ch/~bengio

January 23, 2006

Reminder: Basics on Probabilities

A few basic equalities that are often used:

1. (conditional probabilities)

$$P(A, B) = P(A|B) \cdot P(B)$$

2. (Bayes rule)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

3. If $(\bigcup B_i = \Omega)$ and $\forall i, j \neq i \ (B_i \bigcap B_j = \emptyset)$ then

$$P(A) = \sum_i P(A, B_i)$$

What is a Gaussian Mixture Model?

- A Gaussian Mixture Model (GMM) is a distribution
- The likelihood given a Gaussian distribution is

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

  where d is the dimension of x, μ is the mean and Σ is the covariance matrix of the Gaussian. Σ is often diagonal.
- The likelihood given a GMM is

$$p(x) = \sum_{i=1}^{N} w_i \cdot \mathcal{N}(x|\mu_i,\Sigma_i)$$

  where N is the number of Gaussians and w_i is the weight of Gaussian i, with

$$\sum_i w_i = 1 \text{ and } \forall i : w_i \geq 0$$

Characteristics of a GMM

- While ANNs are universal approximators of functions,
- GMMs are universal approximators of densities.

  (as long as there are enough Gaussians of course)

- Even diagonal GMMs are universal approximators.
- Full rank GMMs are not easy to handle: number of parameters is the square of the number of dimensions.
- GMMs can be trained by maximum likelihood using an efficient algorithm: Expectation-Maximization.

Practical Applications using GMMs

- Biometric person authentication (using voice, face, handwriting, etc):
  - one GMM for the client
  - one GMM for all the others
  - Bayes decision $\Longrightarrow$ likelihood ratio
- Any highly imbalanced classification task
  - one GMM per class, tuned by maximum likelihood
  - Bayes decision $\Longrightarrow$ likelihood ratio
- Dimensionality reduction
- Quantization

Basics of Expectation-Maximization

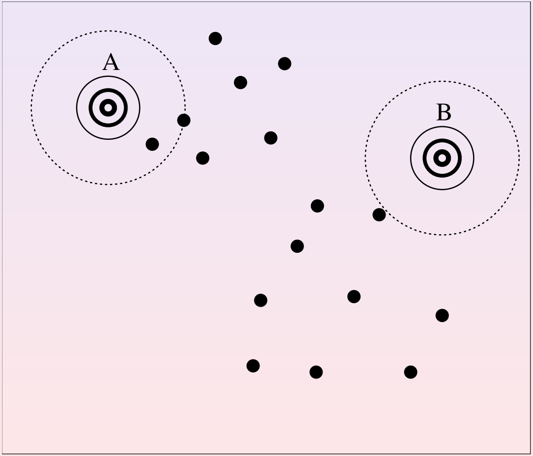
- Objective: maximize the likelihood $p(X|\theta)$ of the data X drawn from an unknown distribution, given the model parameterized by θ:

$$\theta^* = \arg\max_\theta p(X|\theta) = \arg\max_\theta \prod_{p=1}^{n} p(x_p|\theta)$$

- Basic ideas of EM:
  - Introduce a hidden variable such that *its knowledge would simplify the maximization of* $p(X|\theta)$
  - At each iteration of the algorithm:
    - E-Step: estimate the distribution of the hidden variable given the data and the current value of the parameters
    - M-Step: modify the parameters in order to maximize the joint distribution of the data and the hidden variable

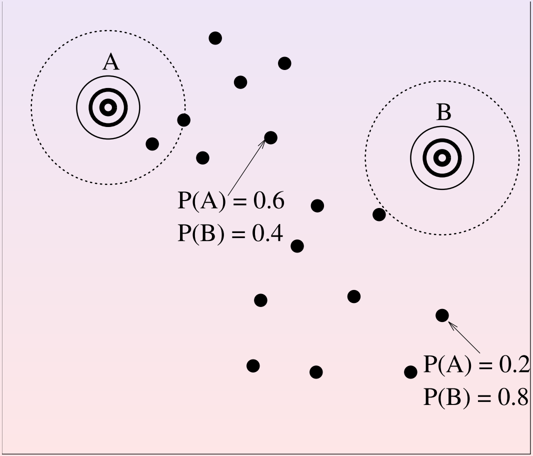
EM for GMM (Graphical View, 1)

Hidden variable: for each point, which Gaussian generated it?



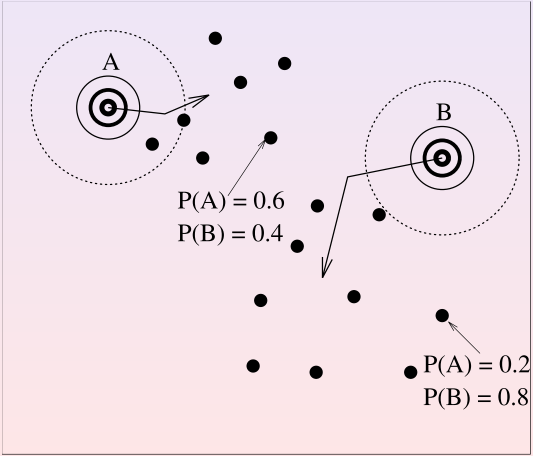
EM for GMM (Graphical View, 2)

E-Step: for each point, estimate the probability that each Gaussian generated it



EM for GMM (Graphical View, 3)

M-Step: modify the parameters according to the hidden variable to maximize the likelihood of the data (and the hidden variable)



EM: More Formally

- Let us call the hidden variable Q.
- Let us consider the following auxiliary function:

$$A(\theta, \theta^s) = E_Q[\log p(X, Q|\theta)|X, \theta^s]$$

- It can be shown that maximizing A

$$\theta^{s+1} = \arg\max_{\theta} A(\theta, \theta^s)$$

always increases the likelihood of the data $p(X|\theta^{s+1})$, and a maximum of A corresponds to a maximum of the likelihood.

EM: Proof of Convergence

First let us develop the auxiliary function:

$$\begin{aligned}
A(\theta, \theta^s) &= E_Q[\log p(X, Q|\theta)|X, \theta^s] \\
&= \sum_{q \in Q} P(q|X, \theta^s) \log p(X, q|\theta) \\
&= \sum_{q \in Q} P(q|X, \theta^s) \log(P(q|X, \theta) \cdot p(X|\theta)) \\
&= \left[\sum_{q \in Q} P(q|X, \theta^s) \log P(q|X, \theta)\right] + \log p(X|\theta)
\end{aligned}$$

EM: Proof of Convergence

- then if we evaluate it at θ^s

$$A(\theta^s, \theta^s) = \left[\sum_{q \in Q} P(q|X, \theta^s) \log P(q|X, \theta^s) \right] + \log p(X|\theta^s)$$

- the difference between two consecutive log likelihoods of the data can be written as

$$\log p(X|\theta) - \log p(X|\theta^s) = A(\theta, \theta^s) - A(\theta^s, \theta^s) + \sum_{q \in Q} P(q|X, \theta^s) \log \frac{P(q|X, \theta^s)}{P(q|X, \theta)}$$

EM: Proof of Convergence

- hence,
  - since the last part of the equation is a Kullback-Leibler divergence which is always positive or null,
  - if A increases, the log likelihood of the data also increases
  - Moreover, one can show that when A is maximum, the likelihood of the data is also at a maximum.

EM for GMM: Hidden Variable

- For GMM, the hidden variable Q will describe which Gaussian generated each example.
- If Q was observed, then it would be simple to maximize the likelihood of the data: simply estimate the parameters Gaussian by Gaussian
- Moreover, we will see that we can easily estimate Q
- Let us first write the mixture of Gaussian model for one x_i:

$$p(x_i|\theta) = \sum_{j=1}^{N} P(j|\theta) p(x_i|j, \theta)$$

- Let us now introduce the following indicator variable:

$$q_{i,j} = \begin{cases} 1 & \text{if Gaussian } j \text{ emitted } x_i \\ 0 & \text{otherwise} \end{cases}$$

EM for GMM: Auxiliary Function

- We can now write the joint likelihood of all the X and q:

$$p(X, Q|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{N} P(j|\theta)^{q_{i,j}} p(x_i|j, \theta)^{q_{i,j}}$$

- which in log gives

$$\log p(X, Q|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{N} q_{i,j} \log P(j|\theta) + q_{i,j} \log p(x_i|j, \theta)$$

EM for GMM: Auxiliary Function

Let us now write the corresponding auxiliary function:

$$
\begin{aligned}
A(\theta, \theta^s) &= E_Q[\log p(X, Q|\theta)|X, \theta^s] \\
&= E_Q\left[\sum_{i=1}^{n}\sum_{j=1}^{N} q_{i,j} \log P(j|\theta) + q_{i,j} \log p(x_i|j, \theta)|X, \theta^s\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{N} E_Q[q_{i,j}|X, \theta^s] \log P(j|\theta) + E_Q[q_{i,j}|X, \theta^s] \log p(x_i|j, \theta)
\end{aligned}
$$

EM for GMM: E-Step and M-Step

$$A(\theta, \theta^s) = \sum_{i=1}^{n} \sum_{j=1}^{N} E_Q[q_{i,j}|X, \theta^s] \log P(j|\theta) + E_Q[q_{i,j}|X, \theta^s] \log p(x_i|j, \theta)$$

- Hence, the E-Step estimates the posterior:

$$\begin{aligned} E_Q[q_{i,j}|X, \theta^s] &= 1 \cdot P(q_{i,j} = 1|X, \theta^s) + 0 \cdot P(q_{i,j} = 0|X, \theta^s) \\ &= P(j|x_i, \theta^s) = \frac{p(x_i|j, \theta^s) P(j|\theta^s)}{p(x_i|\theta^s)} \end{aligned}$$

- and the M-step finds the parameters θ that maximizes A, hence searching for

$$\frac{\partial A}{\partial \theta} = 0$$

  for each parameter (μ_j, variances σ_j^2, and weights w_j).
- Note however that w_j should sum to 1.

EM for GMM: M-Step for Means

$$A(\theta, \theta^s) = \sum_{i=1}^{n} \sum_{j=1}^{N} E_Q[q_{i,j}|X, \theta^s] \log P(j|\theta) + E_Q[q_{i,j}|X, \theta^s] \log p(x_i|j, \theta)$$

$$\begin{aligned}
\frac{\partial A}{\partial \mu_j} &= \sum_{i=1}^{n} \frac{\partial A}{\partial \log p(x_i|j, \theta)} \frac{\partial \log p(x_i|j, \theta)}{\partial \mu_j} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \frac{\partial \log p(x_i|j, \theta)}{\partial \mu_j} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{(x_i - \mu_j)}{\sigma_j^2} = 0
\end{aligned}$$

EM for GMM: M-Step for Means

$$\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{(x_i - \mu_j)}{\sigma_j^2} = 0$$

$\Longrightarrow$ (removing constant terms in the sum)

$$\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot x_i \quad - \quad \sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \mu_j = 0$$

$$\frac{\displaystyle\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot x_i}{\displaystyle\sum_{i=1}^{n} P(j|x_i, \theta^s)} \quad = \quad \hat{\mu}_j$$

EM for GMM: M-Step for Variances

$$A(\theta, \theta^s) = \sum_{i=1}^{n} \sum_{j=1}^{N} E_Q[q_{i,j}|X, \theta^s] \log P(j|\theta) + E_Q[q_{i,j}|X, \theta^s] \log p(x_i|j, \theta)$$

$$\begin{aligned}
\frac{\partial A}{\partial \sigma_j^2} &= \sum_{i=1}^{n} \frac{\partial A}{\partial \log p(x_i|j, \theta)} \frac{\partial \log p(x_i|j, \theta)}{\partial \sigma_j^2} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \frac{\partial \log p(x_i|j, \theta)}{\partial \sigma_j^2} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \left(\frac{(x_i - \mu_j)^2}{2\sigma_j^4} - \frac{1}{2\sigma_j^2} \right) = 0
\end{aligned}$$

EM for GMM: M-Step for Variances

$$\begin{aligned}
\sum_{i=1}^{n} P(j|x_i,\theta^s)\cdot\left(\frac{(x_i-\mu_j)^2}{2\sigma_j^4}-\frac{1}{2\sigma_j^2}\right) &= 0 \\
\sum_{i=1}^{n}\frac{P(j|x_i,\theta^s)(x_i-\mu_j)^2}{2\sigma_j^4}-\sum_{i=1}^{n}\frac{P(j|x_i,\theta^s)}{2\sigma_j^2} &= 0 \\
\sum_{i=1}^{n}\frac{P(j|x_i,\theta^s)(x_i-\mu_j)^2}{\sigma_j^2}-\sum_{i=1}^{n}P(j|x_i,\theta^s) &= 0 \\
\frac{\displaystyle\sum_{i=1}^{n}P(j|x_i,\theta^s)(x_i-\hat{\mu}_j)^2}{\displaystyle\sum_{i=1}^{n}P(j|x_i,\theta^s)} &= \hat{\sigma}_j^2
\end{aligned}$$

EM for GMM: M-Step for Weights

We have the constraint that all weights w_j should be positive and sum to 1:

$$\sum_{j=1}^{N} w_j = 1$$

Incorporating it into the system:

$$J(\theta, \theta^s) = A(\theta, \theta^s) + (1 - \sum_{j=1}^{N} w_j) \cdot \lambda_j$$

where λ_j are Lagrange multipliers.
So we need to derive J with respect to w_j and to set it to 0.

EM for GMM: M-Step for Weights

$$
\begin{aligned}
\frac{\partial J}{\partial w_j} &= \frac{\partial J}{\partial A(\theta,\theta^s)} \frac{\partial A(\theta,\theta^s)}{\partial w_j} - \lambda_j \\
&= 1 \cdot \left(\sum_{i=1}^{n} P(j|x_i,\theta^s) \cdot \frac{1}{w_j} \right) - \lambda_j = 0 \\
\hat{w}_j &= \frac{\sum_{i=1}^{n} P(j|x_i,\theta^s)}{\lambda_j}
\end{aligned}
$$

and incorporating the probabilistic constraint, we get

$$
\hat{w}_j = \frac{\sum_{i=1}^{n} P(j|x_i,\theta^s)}{\sum_{k=1}^{N}\sum_{i=1}^{n} P(k|x_i,\theta^s)} = \frac{1}{n}\sum_{i=1}^{n} P(j|x_i,\theta^s)
$$

EM for GMM: Update Rules

Means $$\hat{\mu}_j = \frac{\displaystyle\sum_{i=1}^{n} x_i \cdot P(j|x_i, \theta^s)}{\displaystyle\sum_{i=1}^{n} P(j|x_i, \theta^s)}$$

Variances $$\hat{\sigma}_j^2 = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \hat{\mu}_j)^2 \cdot P(j|x_i, \theta^s)}{\displaystyle\sum_{i=1}^{n} P(j|x_i, \theta^s)}$$

Weights: $$\hat{w}_j = \frac{1}{n} \sum_{i=1}^{n} P(j|x_i, \theta^s)$$

Initialization

- EM is an iterative procedure that is very sensitive to initial conditions!
- Start from trash $\rightarrow$ end up with trash.
- Hence, we need a good and fast initialization procedure.
- Often used: K-Means.
- Other options: hierarchical K-Means, Gaussian splitting.

Capacity Control

- How to control the capacity with GMMs?
  - selecting the number of Gaussians
  - constraining the value of the variances to be far from 0 (small variances $\Longrightarrow$ large capacity)
- Use cross-validation on the desired criterion (Maximum Likelihood, classification...)

Adaptation Techniques

- In some cases, you have access to only a few examples coming from the target distribution...
- ... but many coming from a nearby distribution!
- How can we profit from the big nearby dataset???
- Solution: use adaptation techniques.
- The most well known and used for GMMs: the Maximum A Posteriori adaptation.

MAP Adaptation

- Normal maximum likelihood training for a dataset X:

$$\theta^* = \arg\max_\theta p(X|\theta)$$

- Maximum A Posteriori (MAP) training:

$$\begin{aligned}\theta^* &= \arg\max_\theta p(\theta|X) \\ &= \arg\max_\theta \frac{p(X|\theta)P(\theta)}{p(X)} \\ &= \arg\max_\theta p(X|\theta)p(\theta)\end{aligned}$$

where $p(\theta)$ represents your prior belief about the distribution of the parameters θ.

Implementation

- Which kind of prior distribution for $p(\theta)$?
- Two objectives:
  - constraining θ to reasonable values
  - keep the EM algorithm tractable
- Use conjugate priors:
  - Dirichlet distribution for weights
  - Gaussian densities for means and variances

What is a Conjugate Prior?

- A conjugate prior is chosen such that the corresponding posterior belongs to the same functional family as the prior.
- So we would like that $p(X|\theta)p(\theta)$ is distributed according to the same family as $p(\theta)$ and tractable.
- Example:
    - Likelihood is Gaussian: $p(X|\theta) = K_1 \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)$
    - Prior is Gaussian: $p(\theta) = K_2 \exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right)$
    - Posterior is Gaussian:

$$
\begin{aligned}
p(X|\theta)p(\theta) &= K_1 K_2 \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) \\
&= K_3 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

Conjugate Prior of Multinomials

- Multinomial distribution:

$$P(X_1 = x_1, \cdots, X_n = x_n | \theta) = \frac{N!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} \theta_i^{x_i}$$

where x_i are nonnegative integers and $\sum_{i=1}^{n} x_i = N$.

- Dirichlet distribution with parameter u:

$$P(\theta | u) = \frac{1}{Z(u)} \prod_{i=1}^{n} \theta_i^{u_i - 1}$$

where $\theta_1, \cdots, \theta_n \geq 0$ and $\sum_{i=1}^{n} \theta_i = 1$ and $u_1, \cdots, u_n \geq 0$.

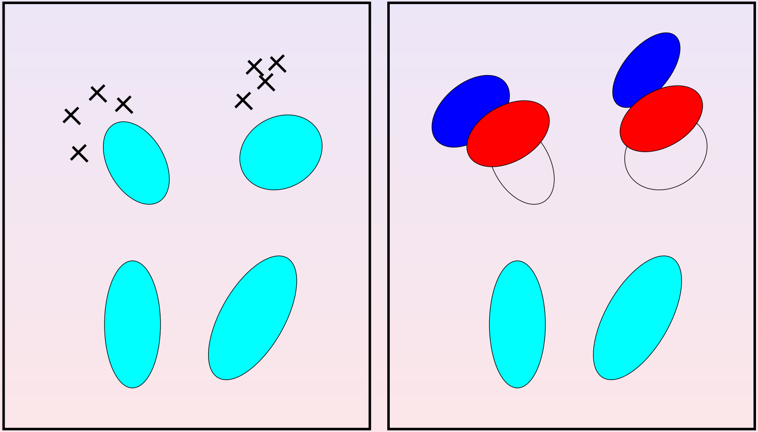
- Conjugate prior = dirichlet with parameter $x + u$:

$$P(X, \theta | u) = \frac{1}{Z} \prod_{i=1}^{n} \theta_i^{x_i + u_i - 1}$$

Examples of Conjugate Priors

likelihood $p(x\|\theta)$	conjugate prior $p(\theta)$	posterior $p(\theta\|x)$
Gaussian(θ, σ)	Gaussian(μ_0, σ_0)	Gaussian(μ_1, σ_1)
Binomial(N, θ)	Beta(r, s)	Beta $(r + n, s + N - n)$
Poisson(θ)	Gamma(r, s)	Gamma$(r + n, s + 1)$
Multinomial$(\theta_1, \cdots, \theta_k)$	Dirichlet $(\alpha_1, \cdots, \alpha_k)$	Dirichlet $(\alpha_1 + n_1, \cdots, \alpha_k + n_k)$

Simple Implementation for MAP-GMMs



Simple Implementation

- Train a generic prior model p with large amount of available data
$$\Longrightarrow \{w_j^p, \mu_j^p, \sigma_j^p\}$$
- One hyper-parameter: $\alpha \in [0, 1]$: faith on prior model
- Weights:
$$\hat{w}_j = \left[\alpha w_j^p + (1-\alpha)\sum_{i=1}^{n} P(j|x_i)\right]\gamma$$
where γ is a normalization factor (so that $\displaystyle\sum_j w_j = 1$)

Simple Implementation

- Means:

$$\hat{\mu}_j = \alpha \mu_j^p + (1-\alpha) \frac{\displaystyle\sum_{i=1}^{n} P(j|x_i) x_i}{\displaystyle\sum_{i=1}^{n} P(j|x_i)}$$

- Variances:

$$\hat{\sigma}_j = \alpha \left(\sigma_j^p + \mu_j^p \mu_j^{p'} \right) + (1-\alpha) \frac{\displaystyle\sum_{i=1}^{n} P(j|x_i) x_i x_i^{'}}{\displaystyle\sum_{i=1}^{n} P(j|x_i)} - \hat{\mu}_j \hat{\mu}_j^{'}$$

Adapted GMMs for Person Authentication

- Person authentication task:

$$\text{accept access if } P(S_i|\mathbf{X}) > P(\bar{S}_i|\mathbf{X})$$

  with S_i a client, $\bar{S}_i$ all the other persons, and $\mathbf{X}$ an access attributed to S_i.
- Using Bayes theorem, this becomes:

$$\frac{p(\mathbf{X}|S_i)}{p(\mathbf{X}|\bar{S}_i)} > \frac{P(\bar{S}_i)}{P(S_i)} = \Delta_{S_i} \approx \Delta$$

- $p(\mathbf{X}|\bar{S}_i)$ is trained on a large dataset
- $p(\mathbf{X}|S_i)$ is MAP adapted from $p(\mathbf{X}|\bar{S}_i)$.
- Δ is found on a separate validation set to optimize a given criterion.