Computational Intelligence: Methods and Applications

Lecture 27
Expectation Maximization algorithm, density modeling

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

General formulation

Given data vectors $D=\{\mathbf{X}^{(i)}\}$, $i=1..n$, and some parametric functions $P(X|\theta)$ that model the density of the data $P(X)$ the best parameters should minimize log-likelihood for all data samples:

$$\theta^* = \arg\min_\theta L(\theta|\mathbf{X}) = -\sum_{i=1}^{n} \ln P\left(\mathbf{X}^{(i)};\theta\right)$$

$P(X|\theta)$ is frequently a Gaussian mixture; for a single Gaussian standard solution will give the formula for mean and variance.

Assume now that X is not complete – features or maybe part of the vector is missing. Let Z=(X,Y) be the complete vector. Joint density:

$$P(\mathbf{Z}|\theta) = P(\mathbf{X},\mathbf{Y}|\theta) = P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)$$

Initial joint density may be formed analyzing cases without missing values; the idea is to maximize the complete data likelihood.

What to expect? E-step.

Original likelihood function $L(\theta|\mathbf{X})$ is based on incomplete information, and since Y is unknown it may be treated as a random variable that should be estimated.
Complete-data likelihood function $L(\theta|\mathbf{Z})=L(\theta|\mathbf{X},\mathbf{Y})$ may be evaluated calculating the expectation of incomplete likelihood over Y. This is done iteratively, starting from initial estimation θ^{i-1} new estimation θ^{i} of parameters and missing values is generated:

$$Q\left(\theta|\theta^{i-1}\right) = E_{\mathbf{Y}}\left[\ln P\left(\mathbf{X},\mathbf{Y}|\theta\right)|\mathbf{X},\theta^{i-1}\right]$$

where X and θ^{i-1} are fixed, θ is a free variable, and the conditional expectation is calculated using the joint distribution of the X, Y variable with fixed X

$$E[Y|X=x] = \int y P_{Y|X}(x,y)dy$$

EM algorithm

First step: calculate expectation over unknown variables;
get the function $Q\left(\theta|\theta^{i-1}\right)$

Second step: maximization, find new values of the parameters:

$$\theta^{i} = \max_\theta Q\left(\theta|\theta^{i-1}\right)$$

Repeat until convergence, $\theta^{i} - \theta^{i-1} < \varepsilon$

EM algorithm converges to local maxima, since during the iterations sequences of likelihoods is monotonically increasing and it is bounded. ET algorithm is sensitive to initial conditions.

Linear combination of *k* Gaussian distributions may be efficiently treated with EM algorithm if one of the hidden variables *v* = 1..*k* that is estimated represents Gaussian number from which data comes.

Example with missing data

4 data vectors, $D = \{X^{(1)}, .. X^{(4)}\}$; $X^T=\{(0,2),(1,0),(2,2),(?,4)\}$, ? = missing

Data model: a Gaussians with diagonal covariance matrix:

$$\theta^{\mathrm{T}} = (\mu_1, \mu_2, \sigma_1, \sigma_2); \quad \theta^{0\mathrm{T}} = (0,0,1,1)$$

Initial value of the parameters are improved calculating expectation over the missing value $y=X_1^{(4)}$; let X_g = known data

$$Q(\theta \mid \theta^0) = E_Y\left[\ln P(\mathbf{X}_g, y \mid \theta) \mid \theta^0, \mathbf{X}_g\right] =$$

$$\int\left(\sum_{i=1}^{3} \ln P(\mathbf{X}^{(i)} \mid \theta) + \ln P\left((y,4)^{\mathrm{T}} \mid \theta\right)\right) P\left(y \mid \theta^0, X_2^{(4)} = 4\right) dy$$

These functions are Gaussians, the first part does not depend on y and the conditional distribution $P(y|x) = P(y,x)/P(x)$

... missing data

Conditional distribution:

$$P\left(y \mid \theta^0; X_2^{(4)} = 4\right) = P\left((y,4)^{\mathrm{T}} \mid \theta^0\right) / P\left(X_2^{(4)} = 4 \mid \theta^0\right)$$

$$= (2\pi)^{-1} \exp\left(-\frac{1}{2}\left(y^2 + 4^2\right)\right) \Big/ \int P\left((y',4)^{\mathrm{T}} \mid \theta^0\right) dy'$$

After some calculation

$$Q(\theta \mid \theta^0) = \sum_{i=1}^{3} \ln P\left(\mathbf{X}^{(i)} \mid \theta\right) - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2} - \ln(2\pi\sigma_1\sigma_2)$$

Maximum of Q gives $\theta^1=(0.75, 2.0, 0.938, 2.0)^{\mathrm{T}}$
EM converges in few iterations here.

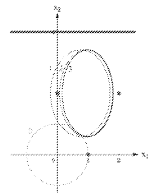

Fig. from Duda, Hart and Stork, Ch. 3.8.

Some applications

- Reconstruction of missing values.
- Reconstruction of images, many medical applications.
- Reconstruction of signals in the presence of noise.
- Unsupervised learning – no information about classes is needed, more than clustering, natural taxonomy.
- Modeling of data, estimation of hidden parameters in mixtures.
- Training of probabilistic models, such as HMM (Hidden Markov models), useful in speech recognition, bioinformatics ...

Associative memory, finding the whole pattern (image) after seeing a fragment – although I have never seen it yet done with EM ...

Book: Geoffrey J. McLachlan, Thriyambakam Krishnan,
The EM Algorithm and Extensions, Wiley 1996

EM demos

Few demonstration of the EM algorithm for Gaussian mixtures may be found in the network.

http://www-cse.ucsd.edu/users/ibayrakt/java/em/

http://www.neurosci.aist.go.jp/~akaho/MixtureEM.html

EM is also a basis for “multiple imputation” approach to missing data. Each missing datum is replaced by m>1 simulated values and m versions of the complete data analyzed by standard methods; results are combined to produce inferential statements that incorporate missing-data uncertainty.

Schafer, JL (1997) Analysis of Incomplete Multivariate Data, Chapman & Hall. Some demo software is available:

http://www.stat.psu.edu/~jls/misoftwa.html

Demonstration of EM in WEKA for clustering data.