Gaussian mixture learning from noisy data

Nikos Vlassis, Jakob Verbeek

Universiteit van Amsterdam

Draft: September 2004

IAS technical report IAS-UVA-04-01

Gaussian mixture learning from noisy data

Nikos Vlassis and Jakob J. Verbeek
Informatics Institute
Faculty of Science
University of Amsterdam
The Netherlands

We address the problem of learning a Gaussian mixture from a set of noisy data points. Each input point has an associated covariance matrix that can be interpreted as the uncertainty by which this point was observed. We derive an EM algorithm that learns a Gaussian mixture that minimizes the Kullback-Leibler divergence to a variable kernel density estimator on the input data. The proposed algorithm performs iterative optimization of a strict bound on the Kullback-Leibler divergence, and is provably convergent.

Keywords: Gaussian mixture, EM algorithm, bound optimization, noisy data.

IAS

intelligent autonomous systems

Contents

Intelligent Autonomous Systems
Informatics Institute, Faculty of Science
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands

Tel (fax): +31 20 525 7461 (7490)
`http://www.science.uva.nl/research/ias/`

Corresponding author:

N. Vlassis
tel: +31 20 525 7522
`vlassis@science.uva.nl`
`http://www.science.uva.nl/~vlassis/`

1 The setup

We are given a set $\{x_1, \ldots, x_n\}$ of measured data points in $\mathbb{R}^d$ together with their uncertainties $\{C_1, \ldots, C_n\}$, where C_i is a $d \times d$ covariance matrix associated with measurement x_i. The task is to learn (fit the parameters of) a k-component Gaussian mixture $p(x)$ from the data, where

$$p(x) = \sum_{s=1}^{k} p(x|s)p(s), \tag{1}$$

and where $p(x|s)$ is the d-variate Gaussian density

$$p(x|s) = \frac{|S_s|^{-1/2}}{(2\pi)^{d/2}} \exp\big[-\frac{1}{2}(x - m_s)^\top S_s^{-1}(x - m_s)\big], \tag{2}$$

parameterized by its mean m_s and covariance matrix S_s. The components of the mixture are indexed by the random variable s that takes values from 1 to k, and $p(s)$ defines a discrete prior distribution over the components. The set of parameters we want to estimate will be denoted $\theta = \{p(s), m_s, S_s\}_{s=1}^k$.

Given that each data point has a covariance associated with it, we can define a variable kernel density estimator [3] in the form

$$f(x) = \frac{1}{n}\sum_{j=1}^{n} f(x|j), \tag{3}$$

where $f(x|j)$ is a d-variate Gaussian like in (2) with known mean x_j and covariance C_j, and where the index j runs over the complete data set.

2 The learning problem

The learning problem can be expressed as an optimization problem where we are looking for the parameter vector that minimizes the Kullback-Leibler divergence between the kernel estimate and the unknown mixture, that is

$$\theta^* = \arg\min_\theta D[f(x)\|p(x;\theta)], \tag{4}$$

where the Kullback-Leibler divergence reads

$$D[f(x)\|p(x)] = \int_x dx f(x) \log\frac{f(x)}{p(x)} = \text{const} - \int_x dx f(x) \log p(x) \tag{5}$$

and where we used the fact that f is fixed. Dropping constants, the objective function to maximize reads

$$L = \sum_{j=1}^{n} \int_x dx f(x|j) \log p(x) \tag{6}$$

which can be regarded as an average log-likelihood function over the kernels $f(x|j)$.

3 An EM approach

In order to maximize L we consider an EM approach in which we iteratively maximize a lower bound of L [1, 2]. This bound F is a function of the current mixture parameters θ and a set of distributions (or 'responsibilities') $q_j(s)$, for $j = 1, \ldots, n$, and it is computed by subtracting from each log-likelihood term a positive quantity (a KL-divergence) as follows:

$$F = \sum_{j=1}^{n} \int_x dx f(x|j) \Big\{ \log p(x) - D[q_j(s) \| p(s|x)] \Big\} \tag{7}$$

which can be rewritten as

$$F = \sum_{j=1}^{n} \sum_{s=1}^{k} q_j(s) \left[\int_x dx f(x|j) \log p(x|s) + \log p(s) - \log q_j(s) \right]. \tag{8}$$

The integral can be analytically computed (dropping constants) as

$$\int_x dx f(x|j) \log p(x|s) = -\frac{1}{2} \big\{ \log |S_s| + (x_j - m_s)^\top S_s^{-1} (x_j - m_s) + \mathrm{Tr}[S_s^{-1} C_j] \big\} \tag{9}$$

where Tr denotes the trace of a matrix.

4 The update equations

We can now directly maximize F w.r.t. the responsibilities $q_j(s)$ and the unknown mixture parameters θ. The resulting update equations are very similar to those of the EM algorithm for noise-free data. For the responsibilities we have

$$q_j(s) \propto p(s|j) \exp \big\{ -\frac{1}{2} \mathrm{Tr}[S_s^{-1} C_j] \big\} \tag{10}$$

where $p(s|j)$ is the Bayes posterior as in the standard EM:

$$p(s|j) = \frac{p(x_j|s) p(s)}{p(x_j)}. \tag{11}$$

For the mixture parameters we get

$$p(s) = \frac{\sum_{j=1}^{n} q_j(s)}{n}, \qquad m_s = \frac{\sum_{j=1}^{n} q_j(s) x_j}{n p(s)}, \qquad S_s = \frac{\sum_{j=1}^{n} q_j(s)(x_j x_j^\top + C_j)}{n p(s)} - m_s m_s^\top. \tag{12}$$

The terms $x_j x_j^\top + C_j$ can be computed in advance. Note that mixture learning with noisy data automatically achieves a sort of regularization: the matrix S_s will always be non-singular as long as the individual matrices C_j are non-singular.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[2] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in graphical models*, pages 355–368. Kluwer Academic Publishers, 1998.

[3] G. Terrell and D. Scott. Variable kernel density estimation. *Ann. Stat.*, 20(3):1236–1265, 1992.

IAS reports

This report is in the series of IAS technical reports. The series editor is Stephan ten Hagen (stephanh@science.uva.nl). Within this series the following titles appeared:

Jelle R. Kok and Nikos Vlassis. *The Pursuit Domain Package.* Technical Report IAS-UVA-03-03, Informatics Institute, University of Amsterdam, The Netherlands, August 2003.

Joris Portegies Zwart, Ben Kröse, and Sjoerd Gelsema. *Aircraft Classification from Estimated Models of Radar Scattering.* Technical Report IAS-UVA-03-02, Informatics Institute, University of Amsterdam, The Netherlands, January 2003.

Joris Portegies Zwart, René van der Heiden, Sjoerd Gelsema, and Frans Groen. *Fast Translation Invariant Classification of HRR Range Profiles in a Zero Phase Representation.* Technical Report IAS-UVA-03-01, Informatics Institute, University of Amsterdam, The Netherlands, January 2003.

M.D. Zaharia, L. Dorst, and T.A. Bouma. *The interface specification and implementation internals of a program.* module for geometric algebra. Technical Report IAS-UVA-02-06, Informatics Institute, University of Amsterdam, The Netherlands, December 2002.

M.D. Zaharia. *Computer graphics from a geometric algebra perspective.* Technical Report IAS-UVA-02-05, Informatics Institute, University of Amsterdam, The Netherlands, August 2002.

J.R. Kok and N. Vlassis. *Mututal modeling of teammate behavior.* Technical Report IAS-UVA-02-04, Informatics Institute, University of Amsterdam, The Netherlands, August 2002.

All IAS technical reports are available for download at the IAS website, http://www.science.uva.nl/research/ias/publications/reports/.