Gaussian mixture model (GMM)

Authors

Alexander Jung[*], Department of Computer Science, Aalto University, Espoo, Finland
[*]Corresponding author: alex.jung@aalto.fi

Abstract

A Gaussian mixture model (GMM) is a probabilistic model for data points with numeric feature vectors. Each data point is generated by first drawing a latent cluster index according to cluster probabilities and then drawing the feature vector from the multivariate normal distribution of that cluster; the marginal distribution is a weighted sum of multivariate normal distributions. The model parameters — cluster probabilities, means, and covariance matrices — are learned by maximum likelihood, typically via the expectation–maximization (EM) algorithm. A fitted GMM delivers soft clustering: the posterior distribution of the latent index grades the membership of each data point in every cluster.

Definition

The nightly minimum temperatures recorded over one year at a weather station do not scatter around a single typical value: cold-season and warm-season nights form two distinct regimes. A GMM is a probabilistic model that captures such multi-regime data: it models the generation of data points with numeric feature vectors $\mathbf{x} \in \mathbb{R}^d$ (Bishop, 2006; Murphy, 2012). It assumes that each data point is generated by first drawing a latent cluster index $I \in \{1, \ldots, k\}$ according to cluster probabilities

$$\mathbb{P}\left(I = c\right) = p_c, \qquad \sum_{c=1}^{k} p_c = 1.$$

Conditioned on $I = c$, the feature vector $\mathbf{x}$ is drawn from a multivariate normal distribution $\mathbb{P}^{(c)} = \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)$. The resulting marginal distribution of $\mathbf{x}$ is therefore a weighted sum of multivariate normal distributions (Fig. 1), i.e.,

$$\mathbb{P} = \sum_{c=1}^{k} p_c \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right).$$

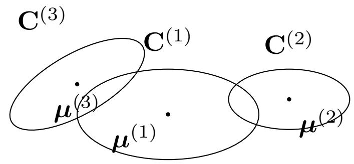

Figure 1: Illustration of a GMM with three components

A GMM is parameterized by the cluster-specific probability p_c, mean $\boldsymbol{\mu}^{(c)}$, and covariance matrix $\mathbf{C}^{(c)}$ for $c = 1, \ldots, k$.

These model parameters are learned from a dataset by maximum likelihood. The resulting optimization problem has no closed-form solution, and the standard method for solving it approximately is the expectation–maximization (EM) algorithm. Fitted by EM to a year of nightly minimum temperatures, a two-component GMM recovers the cold-season and warm-season regimes as its two components. A fitted GMM delivers soft clustering: the posterior distribution $\mathbb{P}\,(I = c \mid \mathbf{x})$ grades the membership of a data point in each cluster instead of assigning it to exactly one — for a night with a minimum temperature between the two regimes, both clusters receive substantial probability.

Cross References

probabilistic model, multivariate normal distribution, clustering, EM, maximum likelihood, soft clustering

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.