Lecture Pattern Analysis

Part 07: Gaussian Mixture Models

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 7, 2021

Introduction

- Gaussian Mixture Models (GMMs) have been covered in Pattern Recognition
- Nevertheless, let's do a quick recap in this lecture[1]
- A GMM models a PDF as sum of K normal distributions $\mathcal{N}$ with weights π_k:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

  Note hereby that $0 \leq \pi_k \leq 1$ and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

  to obtain a proper distribution.
- GMMs are fitted to data with an **Expectation-Maximization** (EM) algorithm

[1] We follow Bishop Sec. 9.2 (including 9.2.1 and 9.2.2).
Hastie/Tibshirani/Friedman Sec. 8.5 and 8.5.1 starts with an instructive 2-component mixture model, but then hastes within only 2 pages through content that is covered in two full sections of Bishop (Sec. 9.2 and 9.3.), so this is probably a little bit too fast.

Preparations for the Probabilistic Model: Hidden Variable z

- We need a K-dim. hidden variable $\mathbf{z}$ to derive the EM algorithm
- Properties of $\mathbf{z}$:
  - $\mathbf{z}$ is a binary indicator vector (“one-hot vector”), i.e.,

$$z_k = \{0, 1\} \tag{3}$$

    and

$$\sum_{k=1}^{K} z_k = 1 \quad , \tag{4}$$

  - The marginal probability of z_k is π_k, i.e., $p(z_k = 1) = \pi_k$, such that

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{5}$$

Joint Distribution over x and z

- The probabilistic modeling of a hidden variable is instructive.
  We consider the joint distribution

  $$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) \cdot p(\mathbf{z}) \tag{6}$$

  consisting of the conditional distribution $p(\mathbf{x}|\mathbf{z})$ and the prior over the hidden variables $p(\mathbf{z})$
- Here, we set

  $$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k} \ , \tag{7}$$

  which results in a single Gaussian component at $z_k = 1$,

  $$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \tag{8}$$

- Insert for the prior $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ as in Eqn. 5 (i.e. $p(z_k = 1) = \pi_k$)

Assembling Everything in the Expectation-Maximization Algorithm

- Marginalization over the hidden variable **z** gives the GMM,

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \qquad (9)$$

- Much ado about nothing? Introduce **z**, cancel it again...?
  Not quite: $p(\mathbf{z})$ drives the Expectation step, $p(\mathbf{x}|\mathbf{z})$ the Maximization step
- EM iteratively fits a GMM to data via Maximum Likelihood:
  1. Initialize $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, Σ
  2. Expectation: determine membership of sample to GMM component
  3. Maximization: optimize GMM components from membership
  4. Goto 2) until convergence
- This is essentially the **soft clustering** variant of k-means

GMM Fitting: Expectation Step (1/2)

- Introduce **responsibilities** $\gamma(z_k)$ that indicate the degree of membership of a sample to a component
- More formally, the responsibility is the likelihood $p(z_k = 1|\mathbf{x})$ that a sample $\mathbf{x}$ belongs to component k:

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) \overset{Bayes}{=} \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \tag{10}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \Sigma_j)} \tag{11}$$

GMM Fitting: Maximization Step (1/2)

- The parameter updates for $\boldsymbol{\mu}_k$, Σ_k, π_k are calculated from maximizing the log likelihood for all samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) \right) \tag{12}$$

- Finding the maximum: set derivatives w.r.t. $\boldsymbol{\mu}_k$, Σ_k, π_k to 0.
- For $\boldsymbol{\mu}_k$:

$$\frac{\partial \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \Sigma_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \tag{13}$$

GMM Fitting: Expectation Step (2/2)

- Setting the derivative to 0 gives

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \cdot \sum_{i=1}^{N} \gamma(z_{ik}) \cdot \mathbf{x}_i \tag{14}$$

with responsibility of component k for all samples N_k, $N_k = \sum_{i=1}^{N} \gamma(z_{nk})$

- The new maxima for Σ_k and $\boldsymbol{\mu}_k$ are found analogously:

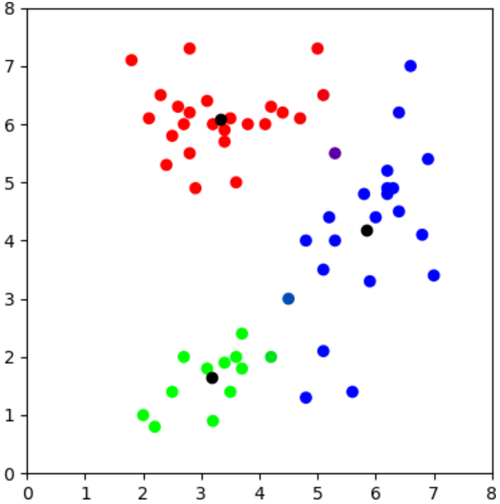
$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \cdot \sum_{i=1}^{N} \gamma(z_{ik}) \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}}) \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}})^{\mathrm{T}} \tag{15}$$

$$\pi_k^{\text{new}} = \frac{N_k}{\sum_{k=1}^{K} N_k} \tag{16}$$

- GMM fitting is locally optimal unless operating on Gaussian distributions

Example Run for $K = 3$

- Black: $\boldsymbol{\mu}_k$ (same starting positions as for k-means)
- Sample chromaticities: Color-coded responsibilities (base colors: red, green, blue)



Example Run for $K = 3$, ML label assignment

- Identical run, but visualization shows component color of maximum responsibility
- Safety notice: only take the maximum on the output. The iteration itself has to use continuous responsibilities