Learning Gaussian Mixtures

Lekshmi Ramesh

Indian Institute of Science
Bangalore

June 16, 2018

Gaussian mixture model

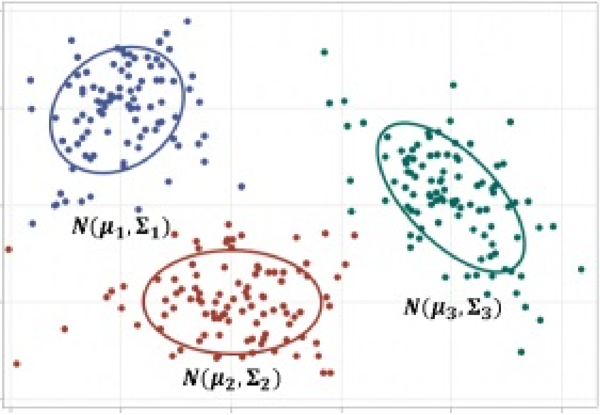
- Gaussian mixture model

$$p(\mathbf{x};\theta) = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_d)$$

  where $\theta = (\{\boldsymbol{\mu}_i\}_{i=1}^k, \{\sigma_i\}_{i=1}^k, \{w_i\}_{i=1}^k)$

- Given data points $\mathbf{x}_1, \ldots, \mathbf{x}_m$, the goal is to fit a mixture of k Gaussians to it

Gaussian mixture model



Separation between components

- We look at the clustering problem, i.e., assigning a label from $\{1, \ldots, k\}$ to each $\mathbf{x}_i$
- Separation between component Gaussians
  - Cannot resolve between two clusters if the means are very close
  - For e.g., in 1-D, if separation is approximately three standard deviations, then clusters are "well separated"
  - What happens in higher dimensions?

Separation between components

- For $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$, expected squared distance from center is

$$\mathbb{E}\|\mathbf{x} - \boldsymbol{\mu}\|_2^2 = d\sigma^2$$

- To resolve between two d-dimensional spherical Gaussians, need separation of $2\sqrt{d}\sigma$ between their centers

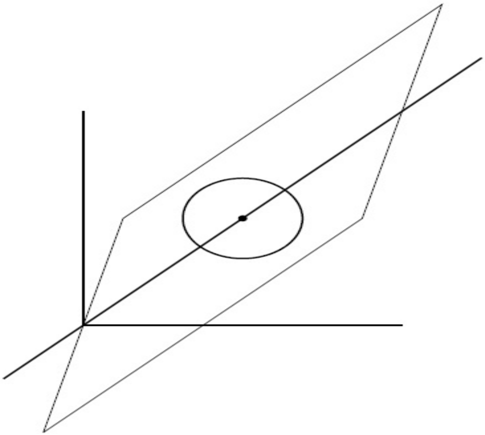
- Separation requirement may be difficult to meet for huge d

- Idea: project data onto a k-dimensional subspace of $\mathbb{R}^d$–now separation requirement would be $\sqrt{k}\sigma$
  This k-dimensional subspace will be the span of the mean vectors

Key ideas

- Project data onto a lower dimension subspace for easier separation condition
- The best fit 1-D subspace to a spherical Gaussian is the line through its center and the origin
- The best fit k-dimensional subspace for k spherical Gaussians is the subspace containing their centers

Best fit subspace to a spherical Gaussian



Projection onto span of means

- Assume that the means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ are known and let U be the subspace spanned by them
- Projecting the data onto U doesn't change the separation, because the means remain unchanged
- Can use distance-based clustering in $\mathbb{R}^k$ which work with $k^{\frac{1}{4}}$ separation
- In reality, we don't know the span of the means
  Find a projection such that the location of the mean vectors is preserved

Span of means and SVD

- Let $V \subset \mathbb{R}^d$ be the span of top k singular vectors of data matrix $X \in \mathbb{R}^{m \times d}$
- If we project rows of X onto V, then separation between mean vectors is approximately preserved: V behaves in the same way as U
- For a vector v, let $\mathrm{proj}_W v$ denote its projection onto subspace W
For a matrix M, let $\mathrm{proj}_W M$ is a matrix whose rows are the rows of M projected onto W

Span of means and SVD

- Facts
  - V is the subspace that maximizes $\|\text{proj}_W A\|$ among all k-dimensional subspaces W
  - U is the subspace that maximizes $\mathbb{E}\|\text{proj}_W A\|$
- Connecting U and V
  - On average, the best k-dimensional subspace approximating X is $U = \text{span}\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$
  - With large probability, the the space spanned by the top k singular vectors of X approximates U well

Concentration result

- For a sufficiently large sample from a mixture of Gaussians, with high probability the subspace found by SVD is very close to the one spanned by the mean vectors
- First show that for an arbitrary subspace W,
  - Show $\mathbb{P}(\|\text{proj}_W A\|^2 > (1+\epsilon)\mathbb{E}\|\text{proj}_W A\|^2) < ke^{\frac{-\epsilon^2 mk}{8}}$
  - Proof uses χ^2 concentration

Main result

- Let the rows of $X \in \mathbb{R}^{m \times d}$ be sampled from a mixture of Gaussians with unifrom weights, means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and variance σ. Let $V \subset \mathbb{R}^d$ be the subspace spanned by the top k singular vectors of X and let U be the span of the means. Then, for $\epsilon \in (0, \frac{1}{2})$, if

$$m \geq \frac{ck}{\epsilon^2} + \left(d \ln \frac{d}{\epsilon} + \frac{d}{d-k} \ln \frac{k}{\delta}\right),$$

  we have w.p. $1 - \delta$

$$\|\mathrm{proj}_U \mathbb{E}X\|^2 - \|\mathrm{proj}_V \mathbb{E}X\|^2 \leq \epsilon m \sigma^2 \left(\frac{d}{k} - 1\right). \qquad (1)$$

- Shows that $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^{'}\|$ is small, where $\boldsymbol{\mu}_i^{'}$ are the projected means

Other approaches

- Algorithms based on random projections
- Algorithms that combine projection idea and EM
- Distribution learning: output a mixture distribution that minimizes a certain loss
- Separation criterion required for all clustering algorithms, not necessary for learning

Reference

S. Vempala and G. Wang. A Spectral Algorithm for Learning Mixture Models. Journal of Computer and System Sciences, 2004.