MIXTURE MODELS & EM

2009年12月18日
下午 05:32

K-means clustering

data set $\{ \mathbf{x}_1, \dots, \mathbf{x}_N \}$ $\quad \mathbf{x}_n \in \mathbb{R}^D$

partion the data set into K clusters

Goal: to find 1) an assignment of data points to clusters and 2) a set of vector $\{\mu_k\}$, such that the sum of the squares of the distances of each data point to its closest vector μ_k, is a minimum.

$\mathbf{x}_n \rightarrow r_{nk} \in \{0, 1\}$, $k = 1, \dots, K$ (1-of-K coding scheme)

$$r_{nk} = \begin{cases} 1, & \mathbf{x}_n \text{ in cluster } K, \\ 0, & \text{otherwise.} \end{cases}$$

$$\underset{\{\mu_k\}\{r_{nk}\}}{\text{minimize}} \quad J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| \mathbf{x}_n - \mu_k \|^2$$

See Fig 9.1

Two-stage optimization

E step $\quad r_{nk}$: $\quad r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \| \mathbf{x}_n - \mu_j \|^2 \\ 0, & \text{otherwise} \end{cases}$

M step $\quad \mu_k$: $\quad \dfrac{\partial J}{\partial \mu_k} = 0 \Rightarrow -2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \mu_k) = 0$

$$\mu_k = \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \Big/ \sum_{n=1}^{N} r_{nk} \quad \text{(cluster mean)}$$

2009年12月18日
下午 05:32

Sequential Update

$$\mu_k^{new} = \mu_k^{old} + \lambda_n \frac{\partial J(\mathbf{x}_n)}{\partial \mu_k}$$

$$= \mu_k^{old} + \lambda_n (-2 r_{nk}) (\mathbf{x}_n - \mu_k^{old})$$

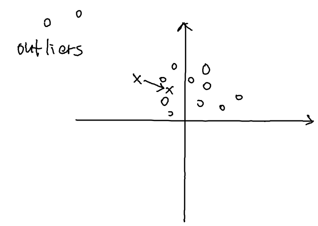
$$= \mu_k^{old} + \eta_{nk} (\mathbf{x}_n - \mu_k^{old})$$

Robustness to Outliers? $\quad l_2$-norm is not robust to outliers

K-medoids

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \mu_k)$$

general dissimilarity measure

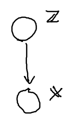


Examples of K-means clustering
image segmentation
vector quantization

Mixtures of Gaussians

2009年12月18日
下午 05:55

Generative Models



$p(z)$

$p(x|z)$

$p(x)$

latent variables $\quad p(z_k = 1) = \pi_k$

$$0 \le \pi_k \le 1 \quad , \quad \sum_{k=1}^{K} \pi_k = 1$$

z uses a 1-of-K coding scheme $\quad \left.\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right\} K$

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

$$p(z_k = 1) = 1 \cdot \ldots \cdot 1 \cdot \pi_k \cdot 1 \cdot \ldots \cdot 1$$

Since $p(x|z_k=1)$ is assumed to be a Gaussian,

$$p(x|z_k=1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

Considering z_k as a selector :

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

the marginal distribution of x

$$p(x) = \sum_{z} p(x, z) = \sum_{z} p(z)\, p(x|z)$$

$$= \sum_{z} \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

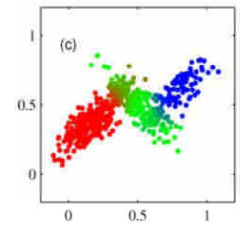
$\Uparrow$

sum over all possible states of $z \in \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \right\}$

2009年12月18日
下午 06:48

responsibility

$$\gamma(z_k) \equiv p(z_k = 1 | x) = \frac{p(z_k=1)\, p(x | z_k = 1)}{\sum_{j=1}^{K} p(z_j=1)\, p(x | z_j = 1)}$$

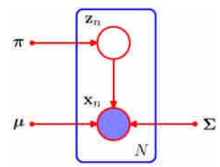


$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

Maximum Likelihood

Given N observations $\{x_1, \ldots, x_N\}$, we want to find μ_k, Σ_k, π_k that maximize the likelihood function

log-likelihood



$$L \equiv \ln p(X | \{\pi_k\}, \{\mu_k\}, \{\Sigma_k\})$$

$$= \ln \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

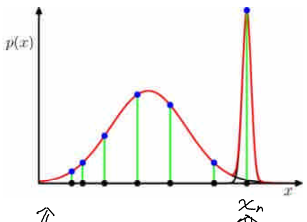
$$= \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

$\Uparrow$

maximize it w.r.t. μ_k, Σ_k, π_k

2009年12月18日
下午 07:01

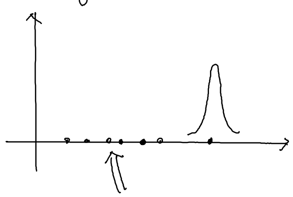
Singularities in the Likelihood Function of Mixtures of Gaussians



$$\mathcal{N}(x_n \mid \mu_j = x_n, \sigma_j^2 I) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_j} \longrightarrow \infty \quad \text{as } \sigma_j \to 0$$

one Gaussian collapses onto a specific data point
the other takes care of the remaining data points

For a single Gaussian model, such a situation would not happen.



Suppose the Gaussian collapes onto a specific data point

these data points have zero likelihoods and thus the overall likelihood goes to zero.

Maximum Likelihood $\quad \frac{\partial L}{\partial \mu_k} = 0, \quad \frac{\partial L}{\partial \Sigma_k} = 0, \quad \frac{\partial L}{\partial \pi_k} = 0$

Recall $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$

2009年12月18日
下午 07:20

① $$0 = \frac{\partial L}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\right\}$$

$$= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\partial \mu_k}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

derivative w.r.t. μ.
(a useful Gaussian identity)

$$\frac{\partial \mathcal{N}}{\partial \mu} = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} \cdot \left(\frac{-1}{2} \cdot 2 \cdot \Sigma^{-1}(x-\mu) \cdot -1\right)$$

$$= \mathcal{N} \, \Sigma^{-1} (x-\mu)$$

Plugged into $\frac{\partial L}{\partial \mu_k}$, we have

$$0 = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} \Sigma_k^{-1} (x_n - \mu_k)$$

multiplying both sides by Σ_k, we obtain

$$\boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \, x_n}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

Next, we set $\frac{\partial L}{\partial \Sigma_k}$ to zero to get Σ_k

(2)

$$0 = \frac{\partial L}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\}$$

$$= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

derivative w.r.t. Σ

$$\frac{\partial \mathcal{N}}{\partial \Sigma} = \frac{1}{(2\pi)^{D/2}} \left(\frac{\partial}{\partial \Sigma} |\Sigma|^{-\frac{1}{2}} \right) \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

$$+ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \frac{\partial}{\partial \Sigma} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

$$= -\frac{1}{2} \Sigma^{-1} \mathcal{N} - \Sigma^{-2} \left(-\frac{1}{2} (x-\mu)(x-\mu)^T \right) \mathcal{N}$$

matrix calculus

$$\frac{\partial}{\partial \Sigma} |\Sigma|^{-\frac{1}{2}} = -\frac{1}{2} |\Sigma|^{-\frac{3}{2}} \frac{\partial}{\partial \Sigma} |\Sigma|$$

$$= -\frac{1}{2} |\Sigma|^{-\frac{3}{2}} |\Sigma| \Sigma^{-1}$$

$$= -\frac{1}{2} |\Sigma|^{-\frac{1}{2}} \Sigma^{-1}$$

$$\frac{\partial}{\partial \Sigma} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\} = \exp\{\ \} \left(-\Sigma^{-2} \right) \left(-\frac{1}{2} (x-\mu)(x-\mu)^T \right)$$

$$\frac{\partial}{\partial A} x^T A x = \frac{\partial}{\partial A} \mathrm{Tr}(x^T A x) = \left(\frac{\partial}{\partial A} x^T A \right) x^T = x x^T$$

$$0 = \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} \left(-\frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \Sigma_k^{-2} (x_n - \mu_k)(x_n - \mu_k)^T \right)$$

multiplying both sides by $2\Sigma_k^2$

$$\boxed{\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}$$

(3)

$$\tilde{L} = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) + \underbrace{\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)}_{\Uparrow \text{ given by the constraint}}$$

$$\frac{\partial \tilde{L}}{\partial \pi_k} = 0$$

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} + \lambda$$

$\Downarrow$ multiplying by π_k, summing over k

$$0 = \sum_{n=1}^{N} 1 + \sum_{k=1}^{K} \pi_k \lambda$$

$$\Downarrow$$

$$\lambda = -N$$

2009年12月18日
下午 08:08

$$\pi_k N = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n \mid \mu_k \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

$$\boxed{\pi_k = \frac{N_k}{N}} \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

SUMMARY

E step:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

M step:

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$