Gaussian mixture regression

STK 802

September 2020

1 The Gaussian distribution / Normal distribution

1.1 The basics

Consider a normally distributed random variable, X, with mean μ and variance σ^2, $X \sim N(\mu, \sigma^2)$

$$\begin{aligned}
f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\
&= N(x|\mu, \sigma^2) \qquad (1) \\
log(f_X(x)) &= -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2 \\
&= -\frac{1}{2} log(2\pi) - \frac{1}{2} log(\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2 \\
&= log\left(N(x|\mu, \sigma^2)\right) \qquad (2)
\end{aligned}$$

1.2 The Gaussian mixture model

A Gaussian mixture distribution consisting of K components can be written as

$$f_X(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2) \qquad (3)$$

with π_k the mixing coefficients and component parameters μ_k and σ_k^2 respectively.

2 Gaussian mixture regression

Gaussian mixture regression is a natural extension of Gaussian mixture modelling. In mixture regression we consider K linear regression models each governed by its own regression parameters $\boldsymbol{\beta}_k$. Considering a mixture of linear regressions using a single target variable y,

$$y_i = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + \epsilon_{i1} & \text{with probability } \pi_1 \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_2 + \epsilon_{i2} & \text{with probability } \pi_2 \\ \dots & \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_K + \epsilon_{iK} & \text{with probability } \pi_K \end{cases} \qquad (4)$$

where

y_i	the i^{th} observation of the response variable
$\boldsymbol{x}_i^T$	the transpose of a p-dimensional vector of explanatory variables, including the intercept term
$\boldsymbol{\beta}_k$	a p-dimensional vector of regression coeffcients of the k^{th} component for i=1,...,K
π_k	are the mixing probabilities $0 < \pi_k < 1$ for all $k = 1, \dots, K$ and $\Sigma_{k=1}^{K} \pi_k = 1$
ϵ_{ik}	random error terms

Note that $\boldsymbol{y} = (y_1, \dots y_n)^T$ a $n \times 1$ vector, $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ . \\ . \\ . \\ \boldsymbol{x}_n^T \end{pmatrix}$ a $n \times p$ matrix and $\boldsymbol{\beta}_k$ a $p \times 1$ vector.

When the component distribution of $y_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)$ for $i = 1, \dots, n$ and $k = 1, \dots, K$ we have a mixture of Gaussian distributions regression model.

The mixture distribution of y therefore is

$$f_Y(y|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k N(y|\boldsymbol{x}^T \boldsymbol{\beta}_k, \sigma^2) \tag{5}$$

with mixing coefficients π_k, conditional means $\boldsymbol{x}^T \boldsymbol{\beta}_k$ and constant variance σ^2. The parameter $\boldsymbol{\theta}$ is the full set of parameters $(\pi_1, \dots, \pi_k; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p; \sigma^2)$.

The log-likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{y}) &= log\, f_Y(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k N(y_i|\boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma^2). \end{aligned} \tag{6}$$

Define a set of binary latent variables, $\boldsymbol{Z} = \{\boldsymbol{z}_i\}$ such that for each observation only one z_{ik} will be 1. That is each observation belongs to only one component.

The complete data log-likelihood function given the observed data $\boldsymbol{y}$ and the latent information $\boldsymbol{Z}$ is

$$l_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}\, log \left\{ \pi_k N(y_i|\boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma^2) \right\}. \tag{7}$$

2.1 Estimation using the EM algorithm

The EM algorithm starts with selecting an initial set of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. In the expectations step these are used to estimate the responsibility of each observation belonging to a specific component, γ_{ik}.

$$\begin{aligned}
\gamma_{ik} &= E(z_{ik}) \\
&= P(z_{ik} = 1 | y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_k) \\
&= P(k | y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_k) \\
&= \frac{P(k\, y_i | \boldsymbol{x}_i, \boldsymbol{\theta}_k)}{P(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}_k)} \\
&= \frac{P(y_i | k, \boldsymbol{x}_i, \boldsymbol{\theta}_k) P(k | \boldsymbol{x}_i, \boldsymbol{\theta}_k)}{P(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}_k)} \\
&= \frac{P(y_i | k, \boldsymbol{x}_i, \boldsymbol{\theta}_k) P(k | \boldsymbol{x}_i, \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} P(y | j, \boldsymbol{x}_i, \boldsymbol{\theta}_k) P(j | \boldsymbol{x}_i, \boldsymbol{\theta}_k)} \\
&= \frac{\pi_k N(y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma^2)}{\sum_{j=1}^{K} \pi_j N(y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \sigma^2)}. \qquad (8)
\end{aligned}$$

Using equation 7 and substituting z_{ik} with the expectation $E(z_{ik}) = \gamma_{ik}$ as in Equation 8 gives

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= E_{\boldsymbol{Z}} l_c(\boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{Z}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left\{ log\, \pi_k + log N(y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma^2) \right\}. \qquad (9)
\end{aligned}$$

In the maximisation step the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ function is maximised with respect to the unknown parameter set $\boldsymbol{\theta}$.

Updating π_k

Maximising $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to π_k taking the constraint $\sum_{k=1}^{K} \pi_k = 1$ into consideration requires the Lagrange multipliers. That is maximising

$$Q^*(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left\{ log\, \pi_k + log N(y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma^2) \right\} + \lambda(\Sigma_{k=1}^{K} \pi_k - 1). \qquad (10)$$

Differentiating $Q^*(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to π_k and λ respectively and setting equal to zero yields

$$\frac{\partial Q^*}{\partial \pi_k} = \sum_{i=1}^{n} \frac{\gamma_{ik}}{\pi_k} + \lambda = 0. \qquad (11)$$

$$\frac{\partial Q^*}{\partial \lambda} = \Sigma_{k=1}^{K} \pi_k - 1 = 0. \tag{12}$$

Summing Equation 11 over k and multiplying by π_k gives

$$\begin{aligned} \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik} + \lambda \sum_{k=1}^{K}\pi_k &= 0 \\ n + \lambda &= 0 \\ \lambda &= -n \end{aligned} \tag{13}$$

since $\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik} = n$ and $\sum_{k=1}^{K}\pi_k = 1$. Solving for π_k by substituting Equation 13 into Equation 11, yields

$$\begin{aligned} \frac{\partial Q}{\partial \pi_k} = \sum_{i=1}^{n}\frac{\gamma_{ik}}{\pi_k} - n &= 0 \\ \pi_k &= \frac{\sum_{i=1}^{n}\gamma_{ik}}{n} \\ &= \frac{n_k}{n} \end{aligned} \tag{14}$$

with $n_k = \sum_{i=1}^{n}\gamma_{ik}$.

Updating $\boldsymbol{\beta}$

Consider only the terms that contains the parameter $\boldsymbol{\beta}_k$ in $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ gives

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{i=1}^{n}\gamma_{ik}\left\{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_k}{\sigma}\right)^2\right\} + const. \tag{15}$$

Partial differentiation of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\beta}_k$ yields

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\beta}_k} = \sum_{i=1}^{n}\gamma_{ik}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_k}{\sigma}\right)\boldsymbol{x}_i^T &= 0 \\ \sum_{i=1}^{n}\gamma_{ik}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_k\right)\boldsymbol{x}_i^T &= 0, \end{aligned} \tag{16}$$

or in matrix notation

$$
\begin{aligned}
\boldsymbol{X}^T \boldsymbol{W_k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k) &= 0 \\
\boldsymbol{X}^T \boldsymbol{W_k}\boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{W_k}\boldsymbol{X}\boldsymbol{\beta}_k &= 0 \\
\boldsymbol{X}^T \boldsymbol{W_k}\boldsymbol{X}\boldsymbol{\beta}_k &= \boldsymbol{X}^T \boldsymbol{W_k}\boldsymbol{y} \\
\boldsymbol{\beta}_k &= \left(\boldsymbol{X}^T \boldsymbol{W_k}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{y}. \qquad (17)
\end{aligned}
$$

with $\boldsymbol{W_k} = diag(\gamma_{ik})$, a $n \times n$ matrix.

Updating σ^2

Consider only the terms that contains the parameter σ^2 in $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ gives

$$
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik} \left\{ -\frac{1}{2} log\sigma^2 - \frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k}{\sigma}\right)^2 \right\} + const. \qquad (18)
$$

Partial differentiation of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to σ^2 yields

$$
\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\sigma^2}} = \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}\gamma_{ik}\frac{1}{\sigma^2} + \frac{1}{2}\gamma_{ik}\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k\right)^2 \frac{1}{\sigma^4} &= 0 \\
\sum_{i=1}^{n}\sum_{k=1}^{K} -\gamma_{ik} + \gamma_{ik}\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k\right)^2 \frac{1}{\sigma^2} &= 0 \\
\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k\right)^2 \frac{1}{\sigma^2} &= \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik} \\
\sigma^2 &= \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k\right)^2}{n} \qquad (19)
\end{aligned}
$$

since $\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik} = n$. The EM algorithm for Gaussian mixture of regressions is given below.

Algorithm 1 Gaussian mixture regression.

1. Choose a set of initial parameters $\boldsymbol{\theta}^{old}$, that is $\pi_1^{old}, \ldots, \pi_k^{old}$, $\beta_1^{old}, \ldots, \beta_k^{old}$ and σ^{2old}
2. In the E-Step, determine the responsibilities

$$\gamma_{ik}^{new} = E(z_{ik}) = \frac{\pi_k N(y_i|\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{old}, \sigma^2)}{\sum_{j=1}^K \pi_j N(y_i|\boldsymbol{x}_i^T \boldsymbol{\beta}_j^{old}, \sigma^2)}.$$

3. In the M-Step update the parameters

$$\begin{aligned}
\pi_k^{new} &= \frac{\sum_{i=1}^n \gamma_{ik}^{new}}{n} = \frac{n_k^{new}}{n}, \\
\boldsymbol{\beta}_k^{new} &= \left(\boldsymbol{X}^T \boldsymbol{W}_k^{new} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}_k^{new} \boldsymbol{y}, \text{ and} \\
\sigma^{2new} &= \frac{\sum_{i=1}^n \sum_{k=1}^K \gamma_{ik}^{new} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k^{new}\right)^2}{n}.
\end{aligned}$$

4. Set $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$
5. Repeat (2) to (4) until convergence.