Mixture of Bernoulli distribution Notes-01

Ye Jiatao

09/05/18

Many open-source ML package contain the mixture model for clustering, but most of them implemented the mixture model based on Gaussian distribution. However, we could face the situation that no all of data are continuous. This code acutally fix the problem when the conditional distribution is bernoulli distribition. The method we have here acutally based on "Bishop: Pattern Recognition and Machine Learning". We added some detailed derivation to clearify the math under the method.

Consider a vector random variable x, the elements inside x are idd and follow Bernoulli distribution prespectively, so that

$$p(\mathbf{x}|\mathbf{u}) = \prod_{i=1}^{D} u_i^{x_i}(1-u_i)^{(1-x_i)}$$

The mean and covariance of this distribution are as follows.

$$E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ ... \\ E[x_D] \end{bmatrix}$$

So that, we get

$$E[x_d] = \sum_{x_d} x_d p(x_d|\mathbf{u}) = u_i$$

where

$$p(x_d|\mathbf{u}) = \sum_{i:i\neq d} \sum_{x_i} p(x_i|\mathbf{u}) = u_i^{x_i}(1-u_i)^{(1-x_i)}$$

A example is as follows.

$$\begin{aligned}
&\sum_{x_2} p(x_1, x_2|u_1 = 0.3, u_2 = 0.6) \\
&= 0.3^{x_1}(1-0.3)^{x_1}0.6^1(1-0.6)^0 + 0.3^{x_1}(1-0.3)^{x_1}0.6^0(1-0.6)^1 \\
&= 0.3^{x_1}(1-0.3)^{x_1}(0.6+0.4) \\
&= 0.3^{x_1}(1-0.3)^{x_1}
\end{aligned}$$

Next, we can also get the covariance of distribution is as follows. Notes that, because the elements of $\mathbf{x}$ are idd, $E[x_i x_j] = E[x_i]E[x_j]$. As a result, the off-diaganol of the final matrix would be all zeros.

$$\begin{aligned}
&cov[\mathbf{x}] \\
&= E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])'] = E[\mathbf{x}\mathbf{x}'] - E[\mathbf{x}]E[\mathbf{x}'] \\
&= \begin{bmatrix} E[x_1^2] & E[x_1x_2] & \dots & E[x_1x_d] \\ E[x_2x_1] & E[x_2^2] & \dots & E[x_2x_d] \\ \vdots & \vdots & \ddots & \vdots \\ E[x_dx_1] & E[x_dx_2] & \dots & E[x_d^2] \end{bmatrix} - \begin{bmatrix} E[x_1]^2 & E[x_1]E[x_2] & \dots & E[x_1]E[x_d] \\ E[x_2]E[x_1] & E[x_2]^2 & \dots & E[x_2]E[x_d] \\ \vdots & \vdots & \ddots & \vdots \\ E[x_d]E[x_1] & E[x_d]E[x_2] & \dots & E[x_d]^2 \end{bmatrix} \\
&= diag(cov[x_1], cov[x_2]...cov[x_d])
\end{aligned}$$

Where the covariance of Bernoulli distribution is $cov[x_i] = u_i(1-u_i)$. Here we should notes that signal multivariate Bernoulli distribution dosen't count for interaction between elements inside $\mathbf{x}$. Next let's consider the mixture of Bernoulli model, given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(z_k)p(\mathbf{x}|z_k)$$

$$p(\mathbf{x}|z_k) = \prod_{i=1}^{D} u_{ki}^{x_i}(1-u_{ki})^{(1-x_i)}$$

Similarly, we can also caculate the expectation and covariance of the mixture distribution.

$$
\begin{aligned}
p(x_d) &= \sum_{i:i\neq d} \sum_{x_i} p(\mathbf{x}) = \sum_{i:i\neq d} \sum_{x_i} \sum_{k=1}^{K} p(z_k) p(\mathbf{x}|z_k) \\
&= \sum_{k=1}^{K} p(z_k) p(x_d|z_k) \\
E[x_d] &= \sum_{x_d} x_d p(x_d) = \sum_{x_d} x_d \sum_{k=1}^{K} p(z_k) p(x_d|z_k) \\
&= \sum_{k=1}^{K} p(z_k) E_{p(x_d|z_k)}[x_d] \\
&= \sum_{k=1}^{K} p(z_k) u_{kd}
\end{aligned}
$$

Thus, we get $E[\mathbf{x}] = \sum p(z_k)\mathbf{u_k}$. Next, we can also derive the cov[x], using the equation $cov[x] = E[xx'] - E[x]E[x]'$.

$$
\begin{aligned}
& cov[\mathbf{x}] = E[\mathbf{x}\mathbf{x}'] - E[\mathbf{x}]E[\mathbf{x}]' \\
& f(\mathbf{x}) = \mathbf{x}\mathbf{x}' \\
& p(\mathbf{x}) = \sum_{k=1}^{K} p(z_k) p(\mathbf{x}|z_k) \\
& E[\mathbf{x}\mathbf{x}'] = E[f(\mathbf{x})] = \sum_{x} f(\mathbf{x}) p(\mathbf{x}) \\
& = \sum_{x} \sum_{k=1}^{K} f(\mathbf{x}) p(z_k) p(\mathbf{x}|z_k) \\
& = \sum_{k=1}^{K} E_{z_k}[f(\mathbf{x})] \\
& = \sum_{k=1}^{K} p(z_k) E_{z_k}[\mathbf{x}\mathbf{x}']
\end{aligned}
$$

Thus, we get

$$\begin{aligned}
cov[\mathbf{x}] &= E[\mathbf{x}\mathbf{x}'] - E[\mathbf{x}]E[\mathbf{x}] \\
&= \sum_{k=1}^{K} p(z_k) E_{z_k}[\mathbf{x}\mathbf{x}'] - E[\mathbf{x}]E[\mathbf{x}]' \\
&= \sum_{k=1}^{K} p(z_k)(\Sigma_k + \mathbf{u}_k\mathbf{u}_k') - E[\mathbf{x}]E[\mathbf{x}]'
\end{aligned}$$

For last setp in the above derivation, the conditional expectation is just the singal Bernoulli ddistribution case, where $\Sigma_k = cov[x]$. It should be noted that the covariance matirx not long diagonal now, which means we can capture the correlation between the variables.