5.3. Learning dyadic data

Every observation in ordinary sample is univariate or multivariate but there is a case that ordinary sample becomes dyadic sample related to two sets of objects, which causes some modifications of mixture model. *Dyadic data* which is also called co-occurrence data (COD) contains co-occurrent events of objects. It is necessary to obtain statistical models to represent dyadic data and fortunately, finite mixture model is the one. Recall that EM is applied to learn mixture model. Here we focus on EM and mixture model for dyadic data or COD.

Given two finite sets = {*x*1, *x*2,…, *xN*) and = {*y*1, *y*2,…, *yM*) with note that *xi* (s) and *yj* (s) represent -objects and -objects, respectively; exactly, they are names of objects. The numbers of -objects and -objects are =*N* and =*M*, respectively. For example, in information retrieval, *xi* (s) are documents and *yj* (s) are keywords. Hence, *xi* and *yj* are not evaluated as numbers. An observational pair (*xi*, *yj*) is called a *co-occurrence* of *xi* and *yj*. Dyadic data or COD contains these co-occurrences with note that a co-occurrence (*xi*, *yj*) can exist more than one time. So, each co-occurrence (*xi*, *yj*) is indexed by an index *r*. As a result, each co-occurrence is denoted by the triple (*xi*, *yj*, *r*) and we have (Hofmann & Puzicha, 1998, p. 1):

|  |  |
| --- | --- |
|  | (5.1.19) |

Where,

Of course, the size of is . As a convention, *xi*(*r*) and *yj*(*r*) indicate that -object and -object at the *r*th co-occurrence are *xi* and *yj*, respectively. Thus, the triplet (*xi*, *yj*, *r*) can be denoted as (*xi*(*r*), *yj*(*r*), *r*). For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and dyadic data of 4 co-occurrences, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, we observe that *x*1 and *y*1 occur together three times at *r*=1, *r*=2, and *r*=4 where as *x*1 and *y*2 occur together one time at *r*=3. In the first co-occurrence (*x*1, *y*1, 1), the notation *x*1(1) indicate that the -object at this co-occurrence is *x*1. In the third co-occurrence (*x*1, *y*2, 3), the notation *y*2(3) indicate that the -object at this co-occurrence is *y*2.

If each co-occurrence of *xi* and *yj* is associated with a value *z* (Hofmann, Puzicha, & Jordan, Learning from Dyadic Data, 1998, p. 1), the triple (*xi*, *yj*, *r*) becomes the quadruplet (*xi*, *yj*, *z*, *r*) which is called *valued co-occurrence* of *xi* and *yj*. The value *z* is called *associative value*. If *z* is value of a variable *Z* then, *Z* is called associative variable. As a result, the sample is called *valued dyadic data* or valued COD.

|  |  |
| --- | --- |
|  | (5.1.20) |

Where,

As a convention *Z*(*r*) or *z*(*r*) indicates that the associative value at *r*th co-occurrence is *Z*=*z*. Thus, the quadruplet (*xi*, *yj*, *Z*, *r*) can be denoted as (*xi*(*r*), *yj*(*r*), *Z*(*r*), *r*). For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and valued dyadic sample of 4 co-occurrences, = {(*x*1, *y*1, 6, 1), (*x*1, *y*1, 8, 2), (*x*1, *y*2, 7, 3), (*x*1, *y*1, 9, 4)}, we observe that *x*1 and *y*1 occur together three times at *r*=1, *r*=2, and *r*=4 where as *x*1 and *y*2 occur together one time at *r*=3. Moreover, at *r*=1, *r*=2, *r*=3, and *r*=4, associative values are *Z*(1)=6, *Z*(2)=7, *Z*(3)=8, and *Z*(4)=9, respectively. Valued dyadic data is special case of dyadic data. As a convention, dyadic data is default if there is no additional information.

Given fixed *xk*, let be the -partitioned subset of which contains co-occurrences whose -objects are fixed at *xk* (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 1). Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.21) |

Dyadic data is partitioned into subsets .

Given fixed *yl*, let be the -partitioned subset of which contains co-occurrences whose -objects are fixed at *yl*. Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.22) |

Dyadic data is partitioned into subsets .

Given fixed *xk* and fixed *yl*, let be the subset of the which contains co-occurrences whose -objects and -objects are fixed at *xk* and *yl*. Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.23) |

Let *n*(*xi*) and *n*(*yj*) denote the number of *xi* and the number of *yj*, respectively.

|  |  |
| --- | --- |
|  | (5.1.24) |

Let *n*(*xi*, *yj*) denote the number of *xi* and *yj*.

|  |  |
| --- | --- |
|  | (5.1.25) |

Let *n*(*xi*|*yj*) and *n*(*yj*|*xi*) denote the frequency of *xi* given *yj* and the frequency of *yj* given *xi*, respectively.

|  |  |
| --- | --- |
|  | (5.1.26) |

For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and dyadic data of 4 co-occurrences, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, we have = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, = = Ø, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*1, 4)}, = {(*x*1, *y*2, 3)}, = = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*1, 4)}, = {(*x*1, *y*2, 3)}, = = = = Ø, *n*(*x*1) = 1, *n*(*x*2) = *n*(*x*3) = 0, *n*(*y*1) = 3, *n*(*y*2) = 1, *n*(*x*1, *y*1) = 3, *n*(*x*1, *y*2) = 1, *n*(*x*2, *y*1) = *n*(*x*2, *y*2) = *n*(*x*3, *y*1) = *n*(*x*3, *y*2) = 0, *n*(*x*1 | *y*1) = 1, *n*(*x*1 | *y*2) = 1, *n*(*x*2 | *y*1) = *n*(*x*2 | *y*2) = *n*(*x*3 | *y*1) = *n*(*x*3 | *y*2) = 0, *n*(*y*1 | *x*1) = 3/4, *n*(*y*2 | *x*1) = 1/4.

Suppose each co-occurrence (*xi*, *yj*) belongs to a latent variable *C* and *C* has *K* values *ck* (s). These values *ck* (s) are called classes or aspects and thus, mixture model for dyadic data is also called aspect model or latent class model which aims to discover the latent variable *C*. Without loss of generality, let *ck* = *k* where *k* = 1, 2,…, *K*. The random variable *C* has discrete distribution such that every value has an associated probability *αk*. Of course, there are *K* probabilities *αk* (s). There are three kinds of mixture model for dyadic data such as symmetric mixture model (SMM), asymmetric mixture model (AMM), and product-space mixture model (PMM).

The mixture model of dyadic data is called symmetric mixture model (SMM) if *αk* (s) are independent from both *xi* and *yj*. SMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 2):

|  |  |
| --- | --- |
|  | (5.1.27) |

Where *αk* is the probability of aspect *k*. Note, *P*(.) denote probability.

The is the probability of *xi* given aspect *k*.

The is the probability of *yj* given aspect *k*.

This implies that *xi* and *yj* are mutually independent in SMM.

The joint probability of *xi*, *yj*, and *k* is:

The parameter of SMM is Θ = (*αk*, *pi*|*k*, *qj*|*k*)*T* in which there are *K*( + + 1) partial parameters *αk*, *pi*|*k*, and *qj*|*k*. Note,

By applying GEM, given dyadic sample , at the *t*th iteration of GEM, given current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, the conditional expectation *Q*(Θ|Θ(*t*)) is:

|  |  |
| --- | --- |
|  | (5.1.28) |

Where,

|  |  |
| --- | --- |
|  | (5.1.29) |

Because there are three constraints

We use Lagrange duality method to maximize to maximize *Q*(Θ|Θ(*t*)). The Lagrange function *la*(Θ, *λ* | Θ(*t*)) is sum of *Q*(Θ|Θ(*t*)) and these constraints, as follows:

Note, *λ* = (*λ*1, *λ*2, *λ*3)*T* where *λ*1≥0, *λ*2≥0, and *λ*3≥0 are called Lagrange multipliers. Of course, *la*(Θ, *λ* | Θ(*t*)) is function of Θ and *λ*. The next parameters Θ(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) at M-step of some *t*th iteration is solution of the equation formed by setting the first-order partial derivatives of Lagrange function regarding Θ and *λ* to be zero.

The first-order partial derivative of Lagrange function regarding *αk* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over *K* aspects {1, 2,…, *K*}, we have:

This means the next parameters *αk*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.30) |

The first-order partial derivative of Lagrange function regarding *pi*|*k* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *pi*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.31) |

Similarly, the next parameters *qj*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.32) |

The two steps of GEM algorithm for SMM at some *t*th iteration are shown in table 5.1.4.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *xi*, *yj*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, according to equation 5.1.29.  *M-step*:  The next parameter Θ(*t*+1) = (*αk*(*t*+1), *pi*|*k*(*t*+1), *qj*|*k*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.30, equation 5.1.31, and equation 5.1.32. |

**Table 5.1.4.** E-step and M-step of GEM algorithm for SMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the SMM itself. When SMM is applied into soft clustering, dyadic data is clustered according to blocks and each *αk* is coverage ratio of cluster *k* (aspect *k*).

The mixture model of dyadic data is called asymmetric mixture model (AMM) if *αk* (s) are only independent from *xi* or from *yj*. Without loss of generality, given *αk* (s) are only independent from *xi*, AMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 3):

|  |  |
| --- | --- |
|  | (5.1.33) |

The *αk*|*i* is the probability of aspect *k* given *xi*.

Where *pi* is the probability of *xi*.

The *qj*|*k* is the conditional probability of *yj* given aspect *k*. Suppose *yj* is dependent from *xi* given *k*, we have:

Note, *qj*|*i* is the conditional probability of *yj* given *xi*, which is defined as follows:

The joint probability of *xi*, *yj*, and *k* is:

The parameter of AMM is Θ = (*αk*|*i*, *pi*, *qj*|*k*)*T* in which there are *K*( + ) + partial parameters *αk*|*i*, *pi*, and *qj*|*k*. Note,

By applying GEM, given dyadic sample , at the *t*th iteration of GEM, given current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, the conditional expectation *Q*(Θ|Θ(*t*)) is:

|  |  |
| --- | --- |
|  | (5.1.34) |

Where,

|  |  |
| --- | --- |
|  | (5.1.35) |

Because there are three constraints

We use Lagrange duality method to maximize to maximize *Q*(Θ|Θ(*t*)). The Lagrange function *la*(Θ, *λ* | Θ(*t*)) is sum of *Q*(Θ|Θ(*t*)) and these constraints, as follows:

Note, *λ* = (*λ*1, *λ*2, *λ*3)*T* where *λ*1≥0, *λ*2≥0, and *λ*3≥0 are called Lagrange multipliers. Of course, *la*(Θ, *λ* | Θ(*t*)) is function of Θ and *λ*. The next parameters Θ(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) at M-step of some *t*th iteration is solution of the equation formed by setting the first-order partial derivatives of Lagrange function regarding Θ and *λ* to be zero.

The first-order partial derivative of Lagrange function regarding *αk* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over *K* aspects {1, 2,…, *K*}, we have:

This means the next parameters *αk*|*i*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.36) |

The first-order partial derivative of Lagrange function regarding *pi* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *pi*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.37) |

The first-order partial derivative of Lagrange function regarding *qj*|*k* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *qj*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.38) |

The two steps of GEM algorithm for AMM at some *t*th iteration are shown in table 5.1.5.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *xi*, *yj*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*αk*|*i*(*t*), *pi*(*t*), *qj*|*k*(*t*))*T*, according to equation 5.1.35.  *M-step*:  The next parameter Θ(*t*+1) = (*αk*|*i*(*t*+1), *pi*(*t*+1), *qj*|*k*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.36, equation 5.1.37, and equation 5.1.38. |

**Table 5.1.5.** E-step and M-step of GEM algorithm for AMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the AMM itself. When AMM is applied into soft clustering, dyadic data is clustered vertically (horizontally) and each *αk*|*i* is coverage ratio of cluster *k* (aspect *k*) according to *xi*. Soft clustering with AMM is also called one-side clustering.

Product-space mixture model (PMM) is derived from SMM with a minor change that the aspect set {1, 2,…, *K*} is Cartesian product of -aspect set {1, 2,…, } and -aspect set {1, 2,…, }. In other words, the aspect space is still symmetric but is checked (stripped) according to two directions and .

|  |  |
| --- | --- |
|  | (5.1.39) |

For every *k* belongs to {1, 2,…, *K*}, there always exists a respective pair: and . However, for each or each , there are many respective *k*.

|  |  |
| --- | --- |
|  | (5.1.40) |

The sign “” denotes correspondence. PMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 4):

|  |  |
| --- | --- |
|  | (5.1.41) |

As usual, *αk* is the probability of aspect *ck* but is the probability of *xi* given of *k* and is the probability of *yj* given of *k*.

The joint probability of *xi*, *yj*, and *k* is:

The parameter of PMM is Θ = (*αk*, , )*T* in which there are *K* + + partial parameters *αk*, , and . Note,

Learning PMM is like learning SMM and so it is not necessary to duplicate the expansion of *Q*(Θ|Θ(*t*)). The two steps of GEM algorithm for PMM at some *t*th iteration are shown in table 5.1.6.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *E-step*:  The conditional probabilities *P*(*k* | *xi*, *yj*, Θ(*t*)), *P*( | *xi*, *yj*, Θ(*t*)), and *P*( | *xi*, *yj*, Θ(*t*)) are calculated based on current parameter Θ(*t*) = , according to equation 5.1.42, equation 5.1.43, and equation 5.1.44.   |  |  | | --- | --- | |  | (5.1.42) | |  | (5.1.43) | |  | (5.1.44) |   *M-step*:  The next parameter Θ(*t*+1) = , which is the maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.45, equation 5.1.46, and equation 5.1.47.   |  |  | | --- | --- | |  | (5.1.45) | |  | (5.1.46) | |  | (5.1.47) | |

**Table 5.1.6.** E-step and M-step of GEM algorithm for PMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the PMM itself. When PMM is applied into soft clustering, dyadic data is clustered in checked (stripped) and each *αk* is coverage ratio of cluster *k* (aspect *k*) but such cluster *k* corresponds to a pair of cluster and cluster . Soft clustering with PMM is also called two-side clustering.

When is valued dyadic data in which every co-occurrence (*xi*, *yj*) is associated with value *z* from random variable *Z* then, SMM is reformed as follows:

|  |  |
| --- | --- |
|  | (5.1.48) |

AMM is reformed as follows:

|  |  |
| --- | --- |
|  | (5.1.49) |

PMM is reformed as follows:

|  |  |
| --- | --- |
|  | (5.1.50) |

Where *fk*(*Z*|*φk*) is the *k*th PDF of *Z* corresponding to the aspect *k*, in which *φk* is parameter of *fk*(*Z*|*φk*). Of course, the parameter Θ now must include all *φk*. It is possible to consider that

Note, suppose *xi* and *yj* are independent from *Z* given aspect *k*, which is the hint to reform these models.

For example, within SMM, the joint probability of *xi*, *yj*, *Z*, and *k* is:

It is only necessary to estimate *φk* because how to estimate other partial parameters was aforementioned. By reforming the conditional expectation *Q*(Θ|Θ(*t*)), it is easy to find out that the next parameter *φk*(*t*+1) is solution of following equation:

|  |  |
| --- | --- |
|  | (5.1.51) |

Where *P*(*k* | *xi*(*r*), *yj*(*r*), Θ(*t*)) is specified by equation 5.1.29, equation 5.1.35, and equation 5.1.42 for SMM, AMM, and PMM, respectively. Especially, if *fk*(*Z*|*φk*) distributed normally, the next parameter *φk*(*t*+1) = (*μk*(*t*+1), Σ*k*(*t*+1))*T* containing mean *μk* and covariance matrix Σ*k* is calculated as follows:

|  |  |
| --- | --- |
|  | (5.1.52) |

Where *P*(*k* | *xi*(*r*), *yj*(*r*), Θ(*t*)) is specified by equation 5.1.29, equation 5.1.35, and equation 5.1.42 for SMM, AMM, and PMM, respectively. Please refer to equation 5.1.17 and equation 5.1.18 to comprehend equation 5.1.52.

**Example 5.1.2.** Suppose = {*x*1, *x*2} and = {*y*1}, and valued dyadic sample of 4 co-occurrences, = {(*x*1, *y*1, 1, 1), (*x*1, *y*1, 9, 2) }, we will learn SMM given by GEM shown in table 5.1.4. Let *Z* be associative variable which distributes normally with mean – variance *φk* = (*μk*, *σk*2)*T* and is learned by equation 5.1.52. Obviously, we have *Z*(1)=1, *Z*(2)=9, *n*(*x*1, *y*1) = 2, and *n*(*x*2, *y*1) = 0. Suppose the number of aspects is *K*=2. The parameter Θ = (*αk*, *pi*|*k*, *qj*|*k*, *φk*)*T* of SMM is initialized as follows:

|  |
| --- |
|  |

At the 1st iteration, E-step, we have:

At the 1st iteration, M-step, we have:

At the 2nd iteration, E-step, we have:

Note, because the probabilities *P*(*k*=1 | *x*2, *y*1, Θ(2)) and *P*(*k*=2 | *x*2, *y*1, Θ(2)) are arbitrary (0/0), they are assigned to be 0.5.

At the 2nd iteration, M-step, we have:

Therefore, GEM stops at the 2nd iteration with the estimate Θ(2) = Θ(3) = Θ\* = (*αk*\*, *pi*|*k*\*, *qj*|*k*\*, *φk*\*)*T*.

|  |
| --- |
|  |

Similarly, it is easy to learn AMM and PMM ■