512 Homework 1 Programming Task Tips

Wenqi He†

1 PathSim

PathSim is discussed in details in the lecture slides, for more information, please see [2].

2 GNetMine

GNetMine [1] is a graph-based regularization algorithm. The input is a heterogeneous network, with some labeled objects. The goal is to predict labels for all types of the remaining unlabeled objects. First, we introduce several related concepts and notations.

Heterogeneous information network. Given m types of data objects, denoted by $\mathcal{X}_1 = \{x_{11}, \dots, x_{1n_1}\}$, $\dots$, $\mathcal{X}_m = \{x_{m1}, \dots, x_{mn_m}\}$, a graph $G = \langle V, E, W \rangle$ is called a heterogeneous information network if $V = \bigcup_{i=1}^{m} \mathcal{X}_i$ and $m \geq 2$, E is the set of links between any two data objects of V, and W is the set of weight values on the links.

Relation graph. In a heterogeneous information network, a relation graph $\mathcal{G}_{ij}$ can be built corresponding to each type of link relationships between two types of data objects $\mathcal{X}_i$ and $\mathcal{X}_j$, where $i, j \in \{1, \dots, m\}$. Let $\mathbf{R}_{ij}$ be an $n_i \times n_j$ relation matrix corresponding to graph $\mathcal{G}_{ij}$. The element at the p-th row and q-th column of $\mathbf{R}_{ij}$ is denoted as $R_{ij,pq}$, representing the weight on link $\langle x_{ip}, x_{jq} \rangle$. Note here we consider undirected graph, so $\mathbf{R}_{ij} = \mathbf{R}_{ji}^T$. For each relation matrix $\mathbf{R}_{ij}$, we further define a diagonal matrix $\mathbf{D}_{ij}$ of size $n_i \times n_i$. The (p, p)-th element of $\mathbf{D}_{ij}$, denoted as $D_{ij,pp}$, is the sum of the p-th row of $\mathbf{R}_{ij}$.

Class. Suppose the number of classes is K. Then for any object type $\mathcal{X}_i$, $i \in \{1, \dots, m\}$, we try to compute a class indicator matrix $\mathbf{F}_i = [\boldsymbol{f}_i^{(1)}, \dots, \boldsymbol{f}_i^{(K)}] \in \mathbb{R}^{n_i \times K}$, where each $\boldsymbol{f}_i^{(k)} = [f_{i1}^{(k)}, \dots, f_{in_i}^{(k)}]^T$ measures the confidence that each object $x_{ip} \in \mathcal{X}_i$ belongs to class k.

Label. In order to encode label information, we set a vector $\boldsymbol{y}_i^{(k)} = [y_{i1}^{(k)}, \dots, y_{in_i}^{(k)}]^T \in \mathbb{R}^{n_i}$ for each data object type $\mathcal{X}_i$ and each class k such that:

$$y_{ip}^{(k)} = \begin{cases} 1 \; if \; x_{ip} \; is \; labeled \; to \; class \; k \\ 0 \; otherwise \end{cases} \tag{2.1}$$

GNetMine designs a set of one-versus-all classifiers for each class $k \in \{1, \dots, K\}$. That is GNetMine will build K classifiers and assign class k to an object if the k-th classifier outputs the maximum estimated confidence for that object among K classifiers.

In order to achieve the goal, GNetMine tries to minimize the following objective function for each class $k \in \{1, \ldots, K\}$:

$$J(\boldsymbol{f}_1^{(k)}, \ldots \boldsymbol{f}_m^{(k)}) = \sum_{i,j=1}^{m} \lambda_{ij} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} R_{ij,pq} \Big(\frac{1}{\sqrt{D_{ij,pp}}} f_{ip}^{(k)} - \frac{1}{\sqrt{D_{ji,qq}}} f_{jp}^{(k)} \Big)^2 + \sum_{i=1}^{m} \alpha_i (\boldsymbol{f}_i^{(k)} - \boldsymbol{y}_i^{(k)})^T (\boldsymbol{f}_i^{(k)} - \boldsymbol{y}_i^{(k)}) \tag{2.2}$$

where:

(1) $\boldsymbol{f}_i^{(k)}$ is the class indicator vector for each object type $\mathcal{X}_i$ and class k, which measures the confidence that an object belongs to class k.

(2) $\boldsymbol{y}_i^{(k)}$ is the ground truth label vector for each object type $\mathcal{X}_i$ and class k, which encodes the prior knowledge that an object is labeled to class k.

(3) $R_{ij,pq}$ is the (p, q)-th element of the relation matrix $\mathbf{R}_{ij}$ between two object types $\mathcal{X}_i$ and $\mathcal{X}_j$.

(4) $D_{ij,pp}$ is the (p, p)-th element of the diagonal matrix $\mathbf{D}_{ij}$, which is derived from the relation matrix $\mathbf{R}_{ij}$.

(5) $D_{ji,qq}$ is the (q, q)-th element of the diagonal matrix $\mathbf{D}_{ji}$, which is derived from the relation matrix $\mathbf{R}_{ji}$. Note in undirected graph, $\mathbf{R}_{ij} = \mathbf{R}_{ji}^T$.

(6) λ_{ij} and α_i are regularization parameters. λ_{ij} controls how much we take into consideration the relationship between object type $\mathcal{X}_i$ and $\mathcal{X}_j$. α_i measures how much we trust the labels of object type $\mathcal{X}_i$.

The first term in the objective function 2.2 follows the intuition that the estimated confidence measures of two objects should be similar if they are linked together. The second term minimizes the difference between the estimated measures and the labels. The intuition is that the class indicator vectors should be similar to ground truth label vectors.

To implement the GNetMine in our classification task, we suggest you to use iterative algorithm. Before introducing the iterative solution, we define the normalized form of $\mathbf{R}_{ij}$:

$$\mathbf{S}_{ij} = \mathbf{D}_{ij}^{(-1/2)} \mathbf{R}_{ij} \mathbf{D}_{ji}^{(-1/2)} \tag{2.3}$$

Then we could rewrite the objective function 2.2 in the following form:

$$J(\boldsymbol{f}_1^{(k)}, \ldots \boldsymbol{f}_m^{(k)}) = \sum_{i,j=1}^{m} \lambda_{ij} ((\boldsymbol{f}_i^{(k)})^T \boldsymbol{f}_i^{(k)} + (\boldsymbol{f}_j^{(k)})^T \boldsymbol{f}_j^{(k)} - 2(\boldsymbol{f}_i^{(k)})^T \mathbf{S}_{ij} \boldsymbol{f}_j^{(k)}) + \sum_{i=1}^{m} \alpha_i (\boldsymbol{f}_i^{(k)} - \boldsymbol{y}_i^{(k)})^T (\boldsymbol{f}_i^{(k)} - \boldsymbol{y}_i^{(k)}) \tag{2.4}$$

Taking derivative with respect to $\boldsymbol{f}_i^{(k)}$ for each i and set it to 0, the iterative form of the algorithm is as follows using fixed-point iteration.

- Step 0: For $\forall k \in \{1,\ldots,K\}$, $\forall i \in \{1,\ldots,m\}$, initialize confidence estimates $\boldsymbol{f}_i^{(k)}(0) = \boldsymbol{y}_i^{(k)}$ and $t = 0$.
- Step 1: Based on the current $\boldsymbol{f}_i^{(k)}(t)$, compute:

$$\boldsymbol{f}_i^{(k)}(t+1) = \frac{\sum_{j=1,j\neq i}^{m} \lambda_{ij}\mathbf{S}_{ij}\boldsymbol{f}_j^{(k)}(t) + 2\lambda_{ii}\mathbf{S}_{ii}\boldsymbol{f}_i^{(k)}(t) + \alpha_i \boldsymbol{y}_i^{(k)}}{\sum_{j=1,j\neq i}^{m} \lambda_{ij} + 2\lambda_{ii} + \alpha_i} \tag{2.5}$$

  for $\forall k \in \{1,\ldots,K\}$, $\forall i \in \{1,\ldots,m\}$.
- Step 2: Repeat step 1 with $t = t+1$ until convergence, i.e., until $\boldsymbol{f}_i^{(k)*} = \boldsymbol{f}_i^{(k)}(t)$ do not change much for all i.
- Step 3: For each $i \in \{1,\ldots,m\}$, assign the class label to the p-th data object of object type $\mathcal{X}_i$ as $c_{ip} = argmax_{1\leq k\leq K} f_{ip}^{(k)*}$, where $\boldsymbol{f}_i^{(k)*} = [f_{i1}^{(k)}*,\ldots,f_{in_i}^{(k)}*]^T$.

For the convergence analysis, please refer to section 2 of this paper [3], which shows a similar proof of the convergence.

In our experiment, we have a heterogeneous information network with 4 object types: author, conference, paper and term. We have 4 classes representing 4 research areas: database, data mining, information retrieval and artificial intelligence. There are 6 relation matrices: paper-author, author-paper, paper-conference, conference-paper, paper-term and term-paper. For parameter setting, you could just set $\alpha_i = 0.1$, and $\lambda_{ij} = 0.2$, $\forall i, j \in \{1,\ldots,4\}$, which is used in the paper.

References

[1] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer, 2010.

[2] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB11*, 2011.

[3] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency.