8 Student-College Matching

Given m students $\mathcal{S} = \{s_1, \ldots, s_m\}$ and n colleges $\mathcal{C} = \{c_1, \ldots, c_n\}$. We have the raw information $x_i^s \in \mathbb{R}^{d_s}$ for a student $s_i \in \mathcal{S}$, and $x_i^c \in \mathbb{R}^{d_c}$ for a college $c_i \in \mathcal{C}$. The notations d_s and d_c represent the number of features in student-profiles and college-profiles, respectively.

Suppose there is a transformation function $f_i^s : \mathbb{R}^{d_s} \mapsto \mathbb{R}^d$ for $s_i \in \mathcal{S}$, $g_i^s : \mathbb{R}^{d_c} \mapsto \mathbb{R}^d$ for all colleges, which is dependent on s_i. The transformed feature vector for s_i is $z_i^s = (z_{i1}^s, z_{i2}^s, \ldots, z_{id}^s) \in \mathbb{R}^d$ and $z_i^c = (z_{j1}^c, z_{i2}^c, \ldots, z_{id}^c) \in \mathbb{R}^d$ for college $c_i \in \mathcal{C}$.

Let all the transformation functions be linear functions, and assuming that

$$\begin{cases} z_i^s & = f_i^s(x_i^s) = W_i^s x_i^s, \\ z_i^c & = g_i^s(x_i^c) = V_i^s x_i^c, \end{cases}$$

where $W_i^s \in \mathbb{R}^{d \times d_s}$ and $V_i^s \in \mathbb{R}^{d \times d_c}$ are the parameter vectors in f_i^s and g_i^s respectively. The transformation function g_i^s is dependent on s_i, because it's pretty hard to obtain the true preference data (i.e. admission requirements or criteria) from the college side.

We expect to learn these (personalized) transform functions for students, and apply them to calculate the match score between s_i and c_i, i.e.

$$p_{ij} = \langle z_i^s, z_i^c \rangle = \langle W_i^s x_i^s, V_i^s x_i^c \rangle, \forall 1 \leq i \leq m, 1 \leq j \leq n.$$

Based on these scores, the prediction accuracy of the model can be evaluated based on some ranking metrics $\mathcal{M}$ (e.g. MAP, NDCG). Suppose student s_i has a true preference ranking π_i over colleges $\mathcal{C}$. We expect the match scores computed using the transformed functions produce a consistent ranking over $\mathcal{C}$ with π_i. Therefore the objective function in learning could be formulated as

$$(W{*}_i^s, V{*}_i^s) = \operatorname*{argmax}_{W_i^s, V_i^s} \mathcal{M}(\pi(r_i^s), \pi_i),$$

where $r_i^s = (r_{i1}, r_{i2}, \ldots, r_{in})^T \in \mathbb{R}^n$, $\pi(r_i^s)$ sorts colleges in descending order in terms of match scores to s_i.

8.1 Pairwise Ranking Model

Let $W^s \in \mathbb{R}^{d \times d_s}$ and $V^c \in \mathbb{R}^{d \times d_c}$ are the feature transformation matrix for students and colleges. Both are independent on a specific student or certain a college. Therefore, given a ranking list over the colleges from each student, we have all the pairs of colleges based on the student's preference profile. To student $s_i \in \mathcal{S}$, he or she may have a preference rankings π_{s_i} over some of the colleges, saying $K = 5$. Here, we assign each of the colleges in π_{s_i} a favorite level or score. The college in the first place will have score K, the second place will be scored $K-1$, and so on. Those colleges failing to be elected get nothing. Let $c^i_{[k]} \in \mathcal{C}$ be the college receives k preference level in π_{s_i}. Furthermore, denote $\mathbb{P}_{s_i} = \{(c^i_{[j]}, c^i_{[k]}), \forall K \geq j > k \geq 0\}$. A college has an implicit ranking position $K+1$ in π_{s_i} if it does not preferred by s_i.

We create a scoring model using the cosine similarity between students and colleges. Therefore, the similarity between s_i and c_1 could be $r_{i1} = \left[W^s x^s_i\right]^T \left[V^s x^c_1\right]$, and $r_{i2} = \left[W^s x^s_i\right]^T \left[V^s x^c_2\right]$. With some simple manipulation, we redefine $W = \left[W^s\right]^T V^c \in \mathbb{R}^{d_s \times d_c}$ as the product of the original two parametric transformation matrices. The similarity between s_i and c_1 could be rewritten as $r_{i1} = [x^s_i]^T W x^c_i$.

We borrow the idea from RankNet, and create a cross entropy loss function based on all pairs of colleges in students' ranking lists. For the pair of colleges $\left(c_{[j]}, c_{[k]}\right) \in \mathbb{P}_{s_i}$. We define the "true" probability of s_i prefers $c_{[j]}$ to $c_{[k]}$ as ${p^*}^i_{jk} = 1/[1+e^{-|j-k|}]$. The bilinear model makes prediction on s_i's the preferences level difference on $c_{[j]}$ and $c_{[k]}$ as $o^i_{12} = [x^s_i]^T W [x^c_{[j]} - x^c_{[k]}]$, and the corresponding probability as $p^i_{jk} = 1/[1+e^{-o^i_{jk}}]$.

Using the two probabilities, we define the CE loss function

$$L^i_{jk} = -{p^*}^i_{jk} \log p^i_{jk} - [1 - {p^*}^i_{jk}] \log[1 - p^i_{jk}] = \log[1 + e^{o^i_{jk}}] - {p^*}^i_{jk} o^i_{jk},$$

and optimize it based on stochastic gradient descent method. Therefore, we have its gradient w.r.t W

$$\frac{\partial L^i_{jk}}{\partial W} = \Big[\frac{e^{o^i_{jk}}}{1 + e^{o^i_{jk}}} - {p^*}^i_{jk}\Big] \frac{\partial o^i_{jk}}{\partial W} = \big[p^i_{jk} - {p^*}^i_{jk}\big] x^s_i \big[x^c_{[j]} - x^c_{[k]}\big]^T.$$

Given the learning rate $\eta > 0$, we update the parameter matrix W iteratively over randomly selected pair of colleges with different preference level for a specific student, saying s_i

$$W \leftarrow W - \eta \big[p^i_{jk} - {p^*}^i_{jk}\big] x^s_i \big[x^c_{[j]} - x^c_{[k]}\big]^T, \forall j > k \geq 0.$$

To explicit present the two individual transformation matrix W^s and W^c, we can predefine a fix $d > 0$ and decompose the learned parametric matrix W based on SVD.