Working Note on Bilinear Prediction of Relatedness

William H. Press

March 28, 2005

1 The Setup

We have a set of objects $\{A\}$ indexed as A_i, with $i = 1, \dots, I$, and another set of objects $\{B\}$ indexed as B_j, with $j = 1, \dots, J$. The A's and B's are not the same kinds of of objects, and, in general, $I \neq J$.

Each A_i has real-valued "features" indexed by $m = 1, \dots, M$. So, A_{im} is the matrix of the values of these features for all the A's. (Each A has the same set of features, but not, in general, the same values.)

Similarly, each B_j has features indexed by $n = 1, \dots, N$. B_{in} is the matrix of the values of these features for all the B's. In general, $M \neq N$, and there need not be any particular correspondence between features of the A's and features of the B's.

As training data, we are given a "relationship matrix" W_{ij} defined by

$$W_{ij} = \begin{cases} +1 & \text{if } A_i \text{ is "related" to } B_j\text{, or} \\ -1 & \text{if } A_i \text{ is "not related" to } B_j \end{cases} \tag{1}$$

There are no other constraints on W_{ij}. That is, the "relationship" need have no special properties (e.g., one-to-one). Note that W_{ij} is not symmetric; in fact, not even generally square.

The problem is: Given the feature matrices A_{im} and B_{jn}, "predict" whether A_i and B_j are related – that is, predict the value of W_{ij}. We want to "learn" how to do this for the W_{ij} provided as training data, and we will then apply it to additional A's and B's not in the training data.

2 The Approach

The approach is pure linear algebra. The basic idea look for a linear combination of A's features (with values denoted A_{i*}) and a linear combination of B's features (with values denoted B_{j*}) the sign of whose product predicts W_{ij} in some best-fit sense,

$$W_{ij} \approx \text{sign}(A_{i*}B_{j*}) \tag{2}$$

More specifically, we proceed as follows:

1. Standardize the features of both $\{A\}$ and $\{B\}$ to have zero mean and unit variance:

$$\begin{aligned}\widehat{A}_{im} &\equiv \frac{A_{im} - \langle A_{xm}\rangle_x}{\langle (A_{ym} - \langle A_{xm}\rangle_x)^2\rangle_y^{\;1/2}} \\ \widehat{B}_{jn} &\equiv \frac{B_{jn} - \langle B_{xn}\rangle_x}{\langle (B_{yn} - \langle B_{xn}\rangle_x)^2\rangle_y^{\;1/2}}\end{aligned} \qquad (3)$$

2. Define A_{i*} and B_{j*} in terms of unknown coefficients α_m and β_n by

$$\begin{aligned} A_{i*} &\equiv \sum_m \alpha_m \widehat{A}_{im} \qquad \text{with } \sum_m \alpha_m^2 = 1 \\ B_{j*} &\equiv \sum_n \beta_n \widehat{B}_{jn} \qquad \text{with } \sum_n \beta_n^2 = 1 \end{aligned} \qquad (4)$$

3. Solve for optimal α's and β's by maximizing the magnitude of the multilinear (linear in A's, B's, and W's) figure-of-merit function

$$\text{F.M.} = \langle A_{i*} W_{ij} B_{j*}\rangle_{ij} \propto \sum_{ijmn} \alpha_m \widehat{A}_{im} W_{ij} \widehat{B}_{jn} \beta_n \equiv \boldsymbol{\alpha}^T \widehat{\mathbf{A}}^T \mathbf{W} \widehat{\mathbf{B}} \boldsymbol{\beta} \qquad (5)$$

subject to the normalization constraints on the α's and β's. The matrix notation is self-explanatory.

3 Method of Solution

Using Lagrange multipliers to impose the constrants, we want to find the extrema of

$$\boldsymbol{\alpha}^T \widehat{\mathbf{A}}^T \mathbf{W} \widehat{\mathbf{B}} \boldsymbol{\beta} - \tfrac{1}{2}\lambda_A \boldsymbol{\alpha}^T \boldsymbol{\alpha} - \tfrac{1}{2}\lambda_B \boldsymbol{\beta}^T \boldsymbol{\beta} \qquad (6)$$

Taking derivatives with respect to each of the α_m's and β_n's gives this (non-standard) eigenvalue problem:

$$\begin{aligned} \widehat{\mathbf{A}}^T \mathbf{W} \widehat{\mathbf{B}} \boldsymbol{\beta} &= \lambda_A \boldsymbol{\alpha} \\ \widehat{\mathbf{B}}^T \mathbf{W}^T \widehat{\mathbf{A}} \boldsymbol{\alpha} &= \lambda_B \boldsymbol{\beta} \end{aligned} \qquad (7)$$

I know of two ways to solve this eigenproblem, a good way and a bad way. The bad way (given first because it is pedagogically more straightforward) is to divide one of the above equations by its eigenvalue $\lambda_{\{A,B\}}$ and then substitute into the other equation. This gives two uncoupled symmetric eigenproblems in standard form,

$$\begin{aligned} \widehat{\mathbf{A}}^T \mathbf{W} \widehat{\mathbf{B}} \widehat{\mathbf{B}}^T \mathbf{W}^T \widehat{\mathbf{A}} \boldsymbol{\alpha} &= (\lambda_A \lambda_B) \boldsymbol{\alpha} \\ \widehat{\mathbf{B}}^T \mathbf{W}^T \widehat{\mathbf{A}} \widehat{\mathbf{A}}^T \mathbf{W} \widehat{\mathbf{B}} \boldsymbol{\beta} &= (\lambda_A \lambda_B) \boldsymbol{\beta} \end{aligned} \qquad (8)$$

The first equation will have M eigenvalues, all non-negative (since the matrix is a "perfect square"). The second equation will have N non-negative eigenvalues, identical to those of the first equation, except that if $M \neq N$, the larger problem will be padded out with additional zero eigenvalues. (Proof left as exercise for reader.)

Chosing any eigenvalue λ identical between the first and second problem, its corresponding eigenvectors are solutions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that will satisfy the original problem, equation (7), with

$$\lambda_A = \lambda_B = \sqrt{\lambda} \tag{9}$$

What is "bad" about this method is that in effectively squaring the original matrix we have squared the condition number of the problem, so the solution is numerically sensitive to roundoff error.

The "good" solution, which gives identical results but more stably (and with less work) is this:

Compute the singular value decomposition (SVD) of the matrix $\widehat{\mathbf{A}}^T\mathbf{W}\widehat{\mathbf{B}}$, namely

$$\widehat{\mathbf{A}}^T\mathbf{W}\widehat{\mathbf{B}} = \mathbf{U}\mathrm{diag}(\mathbf{w})\mathbf{V}^T \tag{10}$$

where $\mathbf{U}$ and $\mathbf{V}$ are column-orthonormal matrices. Then for each singular value w_k, the corresponding column of $\mathbf{U}$ is a solution $\boldsymbol{\alpha}$ and the corresponding column of $\mathbf{V}$ is the corresponding $\boldsymbol{\beta}$. The values of λ_A and λ_B are both w_k, and we also have (cf. equation 5),

$$\boldsymbol{\alpha}^T\widehat{\mathbf{A}}^T\mathbf{W}\widehat{\mathbf{B}}\boldsymbol{\beta} = w_k \tag{11}$$

Proof of all this left to the reader.

4 Discussion

We started out looking for a single pair of linear combinations A_{i*} and B_{j*} that extremize the figure of merit (5). In the end, we have found (generically) $\min(M, N)$ such pairs, all mutually orthogonal.

The merit of each pair is given by the corresponding singular value w_k, so, in practice, we will only be interested in pairs with large singular values – significantly larger than might occur by chance in some randomized control relationship matrix W'_{ij}, for example.

The pairs are orthogonal in the sense that, if A_{i*} and A'_{i*} are two such extremal solutions,

$$\langle A_{i*}A'_{i*}\rangle_i \approx 0 \tag{12}$$

(meaning exactly zero over the sample, and approximately zero over the population). And similarly for the B_{j*}'s. This suggests that we may be able to use more than one solution in a single prediction of a W_{ij}. For example, we might estimate log-odds for each pair of "eigenfeatures" used separately (presumably these are generally decreasing as the singular values get small), and then sum all the log-odds. I haven't looked into this yet.