Sports Ranking Applications in a Math for Data Science Course

Tom Tegtmeyer

Trinity University, San Antonio, TX

Introduction

At Trinity University, we teach a Math for Data Science Course. It has a Calculus I/Math for Business and Economics (Precalculus/Calc I Lite) prerequisite and is mostly an applied linear algebra course with a few other topics.

There are several ranking methods that are applications of topics in the course including the Colley Matrix Method (systems of equations), the Massey Method (least squares solutions), and the Langville & Meyer and Oracle Methods (Markov chains/eigenvectors).

Running Example

We'll use these results to illustrate the various methods. In these standings, ties are broken by point difference (PD).

Results

Winner	Score	Loser	Score
B	31	A	10
C	21	A	17
A	24	D	20
B	42	E	10
D	10	C	7
E	31	D	10

Standings

Team	W	L	PF	PA	PD
B	2	0	73	20	53
C	1	1	28	27	1
E	1	1	41	52	-11
A	1	2	51	72	-21
D	1	2	40	62	-22

The Colley Matrix Method (Systems of Linear Equations)

Wes Colley's method [3] involves solving a system of equations

$$r_i = \frac{1 + (w_i - l_i)/2 + \sum_{j \in O_i} r_j}{2 + t_i}, i = 1, \dots, n$$

where

- n is the number of teams,
- r_i is Team i's rating,
- w_i is the number of wins for Team i,
- l_i is the number of losses for Team i,
- t_i is the number of games played by Team i, and
- O_i is a list of Team i's opponents, with possible repeats.

Note that $t_i = w_i + l_i$ if there are no ties.

We can write it in matrix form $\mathbf{Cr} = \mathbf{b}$, where $\mathbf{r}$ is the ratings vector,

$$c_{ij} = \begin{cases} 2 + t_i, & i = j \text{ (diagonal entries)} \\ -n_{ij}, & i \neq j \end{cases},$$

with n_{ij} being the number of times Teams i and j have played, and $\mathbf{b}$ is the column vector with entries

$$b_i = 1 + \frac{w_i - l_i}{2}.$$

In our example, the equation is

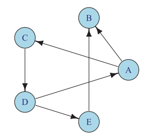
$$\begin{bmatrix} 5 & -1 & -1 & -1 & 0 \\ -1 & 4 & 0 & 0 & -1 \\ -1 & 0 & 4 & -1 & 0 \\ -1 & 0 & -1 & 5 & -1 \\ 0 & -1 & 0 & -1 & 4 \end{bmatrix} \mathbf{r} = \begin{bmatrix} 0.5 \\ 2 \\ 1 \\ 0.5 \\ 1 \end{bmatrix}$$

and the solution is $\mathbf{r} = (0.412, 0.735, 0.447, 0.377, 0.528)$.

Markov Chain/Eigenvector Methods

The Langville & Meyer and Oracle Methods involve Markov chains and eigenvectors. Both methods use tricks to create regular transition matrices. The ratings vectors are the stable eigenvectors guaranteed by the Perron-Frobenius Theorem.

Below left are the results in graph form, with the directed edges pointing from the loser to the winner. Below right is the associated *transposed* adjacency matrix $\mathbf{A}$ so a_{ij} is the number of times team i beat team j.



$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The Langville & Meyer Markov Method

This Markov method [4] converts the adjacency matrix into a column substochastic matrix by assigning equal probability to each directed edge from a vertex. To make it stochastic, if column j has all zeros (representing an undefeated team), we replace it with the jth column of the identity matrix.

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \to \mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 1 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

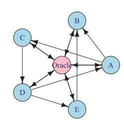
To make the transition matrix regular (indeed, positve), this method, similar to PageRank [2], introduces a teleportation (or damping) matrix $\mathbf{E}$, with $e_{ij} = 1/n$ for all i, j, creating a matrix

$$\mathbf{G} = \alpha\mathbf{S} + (1 - \alpha)\mathbf{E},$$

with a traditional value of $\alpha = 0.85$. The ratings vector is a probability eigenvector $\mathbf{r}$ corresponding to the eigenvalue 1. In practice, we can find $\mathbf{r}$ by raising $\mathbf{G}$ to a high enough power k so that all columns of $\mathbf{G}^k$ agree. For our small schedule, we get a ratings vector of $\mathbf{r} = (0.063, 0.738, 0.057, 0.078, 0.063)$.

The Oracle Method

The Oracle Method [1] induces regularity by introducing a new team, called the Oracle, that both defeats and loses to all other teams. The number of wins and losses involving the Oracle can be customized, depending on the purpose of the rankings. Here, the number of losses to the Oracle is 1, while the number of wins is the number of actual wins plus one. (Note that the number of wins are the row sums in the original adjacency matrix.) If you want to incorporate points or other factors, you can.



$$\mathbf{A}' = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 2 \\ 1 & 0 & 0 & 0 & 1 & 3 \\ 1 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \to \mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & 1/3 & 0 & 2/11 \\ 1/3 & 0 & 0 & 0 & 1/2 & 3/11 \\ 1/3 & 0 & 0 & 0 & 0 & 2/11 \\ 0 & 0 & 1/2 & 0 & 0 & 2/11 \\ 0 & 0 & 0 & 1/3 & 0 & 2/11 \\ 1/3 & 1 & 1/2 & 1/3 & 1/2 & 0 \end{bmatrix}$$

We get a ratings vector of $\mathbf{r} = (0.107, 0.191, 0.103, 0.119, 0.107, 0.371)$, with the last (and highest) rating belonging to the Oracle.

Least Squares Methods: The Massey Method

The Massey Method, developed by Ken Massey [5, 4], uses the difference in ratings to create a point spread. The point spreads in the completed games give us our system of equations. Note that the coefficient matrix has dependent columns.

$$\begin{cases} r_A - r_B & = -21, \\ r_A - r_C & = -4 \\ r_A - r_D & = 4 \\ r_B - r_E & = 32 \\ r_C - r_D & = -3 \\ r_D - r_E & = -21 \end{cases} \to \underbrace{\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} r_A \\ r_B \\ r_C \\ r_D \\ r_E \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} -31 \\ -4 \\ 4 \\ 32 \\ -3 \\ -21 \end{bmatrix}}_{\mathbf{y}}$$

We want to find a least squares solution using the normal equations $\mathbf{X}^T\mathbf{X}\mathbf{r} = \mathbf{X}^T\mathbf{y}$. Note the similarity between the matrix $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and the Colley Matrix, but the columns are dependent. The vector $\mathbf{p} = \mathbf{X}^T\mathbf{y}$ contains the point differences for each team, which is useful for computation.

$$\underbrace{\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & 0 & 0 & -1 \\ -1 & 0 & 2 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ 0 & -1 & 0 & -1 & 2 \end{bmatrix}}_{\mathbf{M}} \mathbf{r} = \underbrace{\begin{bmatrix} -21 \\ 53 \\ 1 \\ -22 \\ -11 \end{bmatrix}}_{\mathbf{p}} \to \underbrace{\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & 0 & 0 & -1 \\ -1 & 0 & 2 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}}_{\mathbf{M}'} \mathbf{r} = \underbrace{\begin{bmatrix} -21 \\ 53 \\ 1 \\ -22 \\ 0 \end{bmatrix}}_{\mathbf{p}'}$$

To get a unique solution*, we replace the bottom row of $\mathbf{M}$ with a row of ones to get $\mathbf{M}'$, and replace the last entry in $\mathbf{p}$ with a zero to get $\mathbf{p}'$. This will make the ratings sum to zero. The solution is the Massey ratings vector $\mathbf{r} = (-5.45, 24.18, -8, -11.55, 0.82)$.

*Note: if the schedule graph is not connected, the matrix $\mathbf{M}'$ could still be singular.

Rankings Summary

The standings look like this.

Standings

Team	W	L	Colley	L & M	Oracle	Massey
A	1	2	4	T3	T3	3
B	2	0	1	1	1	1
C	1	1	3	5	5	4
D	1	2	5	2	2	5
E	1	1	2	T3	T3	2

Projects

I've had students choose to do ranking method projects in the Math for Data Science course. These have involved

1. Getting and cleaning schedule data from websites for sports of their choosing, including NFL, cricket, and e-sports. The level of difficulty can vary, depending on the sport and data source.
2. Writing code to convert schedule data to adjacency and related matrices.
3. Computing the ratings vectors using the appropriate matrix operations.
4. Comparing results from different methods and determining why they are similar or different.

References

[1] Eduardo C. Balreira, Brian K. Miceli, and Thomas Tegtmeyer. An oracle method to predict NFL games. *Journal of Quantitative Analysis in Sports*, 10(2):183–96, 2014.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 33:107–17, 1998.

[3] Wesley N. Colley. Colley's bias free college football ranking method: The Colley matrix explained. https://www.colleyrankings.com/matrate.pdf, 2002.

[4] Amy N. Langville and Carl D. Meyer. *Who's #1? The Science of Rating and Ranking*. Princeton University Press, 2012.

[5] Kenneth Massey. Statistical models applied to the rating of sports teams. Bachelor's thesis, Bluefield College, 1997.