Using March Madness in the first Linear Algebra course

Steve Hilbert

Ithaca College

Hilbert@ithaca.edu

Background

- National meetings
- Tim Chartier
- 1 hr talk and special session on rankings
- Try something new

Why use this application?

- This is an example that many students are aware of and some are interested in.
- Interests a different subgroup of the class than usual applications
- Interests other students (the class can talk about this with their non math friends)
- A problem that students have “intuition” about that can be translated into Mathematical ideas
- Outside grading system and enforcer of deadlines (Brackets “lock” at set time.)

How it fits into Linear Algebra

- Lots of “examples” of ranking in linear algebra texts but not many are realistic to students.
- This was a good way to introduce and work with matrix algebra.
- Using matrix algebra you can easily scale up to work with relatively large systems.

Filling out your bracket

- You have to pick a winner for each game
- You can do this any way you want
- Some people use their “ knowledge”
- I know Duke is better than Florida, or Syracuse lost a lot of games at the end of the season so they will probably lose early in the tournament
- Some people pick their favorite schools, others like the mascots, the uniforms, the team tattoos…

Why rank teams?

- If two teams are going to play a game ,the team with the higher rank (#1 is higher than #2) should win.
- If there are a limited number of openings in a tournament, teams with higher rankings should be chosen over teams with lower rankings.
- Ranking methods may be more objective than some other methods.

What is a Ranking?

- We assign to each team a number
- (in many systems a number between 0 and 1)
- We will call this number the team’s rating
- The team with the highest rating is ranked #1
- The team with the next highest rating is #2 etc

How to rank

- Vote (possibly by “experts”)
- AP poll
- The tournament selection “seeds”
- Use Math to process data and try to come up with a rating that reflects the results of games played before the tournament

First Idea : Winning Percentage

- Calculate # of wins/# of games (this is the winning percentage)
- The team with the highest winning percentage is #1 and so on.
- Easy to do.
- “Should” be related to future performance.
- Problems?
- A team that plays lots of weak teams may have a higher winning percentage than a team that plays lots of strong teams

An Example

Mar 2 Big East	Cinncinatti	Uconn	DePaul	Georgetown
Cinn		L W	W	L
UConn	W L		W W	L
DePaul	L	L L		L
Georgetown	W	W	W	

We will calculate the winning % for each team = #wins/#games

Cinn had 2W and 2 L so its winning percentage was 2/4 = 50% or .5

Conn had 3W and 2 L so its winning percentage was 3/5 = 60% or .6

DePaul had 0W and 4 L so its winning percentage was 0/4 = 0% or .0

Georgetown had 3W and 0 L so its winning percentage was 3/3 = 100% or 1

Results

- So based on this criteria Georgetown would be #1 (the most likely to win)
- UConn would be #2, Cinn would be #3 and
- DePaul would be #4
- And according to these rankings if Georgetown played DePaul then Georgetown should win and if UConn played Cinn , UConn should win.
- Most people would agree that based on this data Georgetown should be strongly favored over DePaul. However if we look at UConn and Cinn it appears that the main reason UConn was ranked above Cinn was that UConn played DePaul twice and Cinn only played DePaul once.

Strength of Schedule

- One method to try to get a “better ranking” is to try to incorporate the idea of “Strength of Schedule”. So instead of “ a win is a win” (all games are equal) we will try to incorporate a method to count a win over a strong team as weighing more than a win over a weak team. So a team that plays lots of strong teams may be ranked higher than a team with a higher winning percentage that plays lots of weak teams.
- Many students in the class brought this idea up.

Vectors for our example

- Cinn=team1,Uconn=team2,DePaul=team3,

Georgetown = team 4 and the first coordinate of each vector corresponds to Cinn and so on.

- w is the wins vector so w = [2,3,0,3]' So Cinn won 2 games, Uconn won 3 games etc
- l is the losses vector so l = [2,2,4,0]'
- Consistency check : total wins should equal total losses since each game has a winner and loser and these games only involved the 4 teams
- t = w+l = [4,5,4,3]’ this gives the total number of games played by each team

Making Things Complicated

- When we calculate the winning percentage we solved 4 equations that were not interconnected $r_i = w_i/t_i$
- In linear algebra a common theme is to find a method to take a problem with interconnected variables and transform it into an equivalent problem where the variables are not connected. (For example Ax=b and rref) However we are looking for a way to relate the rankings

First Tweak
Adjusted winning percentage

- At the start of the season each team's winning percentage would be 0/0 which is undefined. So we will assign a rating to each team which is (1+#wins)/(2+#games) So ri = (1+wi)/(2+ti) and each team starts with a rating of .5, (So all teams start off the same) and since in each game the denominator increases by 1, the losers rating will decrease and the winners rating will increase after each game.
- Since the average of all ratings should be ~.5 The total of the ratings for all the teams played so far by team i should be roughly .5*(number of games) throughout the season.

Bringing in other teams ratings

- Now when we used winning percentage each rating only depends on wins and losses

- $(2+ti)ri = 1+wi$
- or
- $(2+wi+li)\ ri = 1+wi$

- We can replace wi by $(wi/2 + wi/2)$

- And now we add $0 = li/2 - li/2$ and rearrange so we have

- $wi = \frac{1}{2}(wi-li) + \frac{1}{2}(wi+li)$

Next comes the Magic

- The term ½(wi+l1) =½(ti) and we will approximate this by the sum of the ratings of the teams that team i played. (If team i played team j 3 times then we will have rj +rj + rj in the sum.)This idea is the Colley Model
- **This brings the rating of the opponents into the equation and if team i plays strong opponents the sum of rankings will be larger than if they play weak opponents. These rankings are called Colley Rankings.**

Back to the example

- This gives us a set of linear equations for the rankings
- For r1 we have $(2+4)r1 = 1 + \frac{1}{2}(2 - 2) + 2r2 + r3 + r4$
- For r2 we have $(2+5)r2 = 1 + \frac{1}{2}(3 - 2) + 2r1 + 2r3 + r4$
- For r3 we have $(2+4)r3 = 1 + \frac{1}{2}(0 - 4) + r1 + 2r2 + r4$
- For r4 we have $(2+3)r4 = 1 + \frac{1}{2}(3 - 0) + r1 + r2 + r3$

Write this as Ax = b

- This gives us 4 equations in the 4 unknowns r1,r2,r3,r4
- (The Colley Equations)
- Put all the unknowns on the left and we have
-
- 6r1 -2r2 -r3 -r4 =1+0 = 1
-
- -2r1 +7r2 -2r3 -r4 = 1 + ½ (1)=3/2
-
- -r1 –2r2 + 6r3 – r4 = 1 + ½ (0-4) = -3
-
- -r1 –r2 –r3 +5r4 = 1+ ½ (3-0) = 5/2
-

Colley Equations

- In matrix vector form we can write this as A**x** = **b**
-
- With **x**= [r1 , r2, r3, r4]’, **b** = [1, 3/2, -3, 5/2]’ and the coefficient matrix
-
- A=   6   -2   -1   -1
-
-        -2   7   -2   -1
-
-        -1   -2   6   -1
-
-        -1   -1   -1   5

How to Construct A

- The entry in row i and column j is the negative of the # of times team i played team j. So the entry in row j and column i should be the same number since the number of times team j played team i is the same as the number of times team i played team j. So the matrix A is symmetric.

- In order to construct the matrix we only need to know how many times team i played team j to get the i,j entry for $i \neq j$ and the total number of games played by team 1 to get the diagonal entry i,i . (the i,i entries are 2 +the number of games played by team i.)

Colley Rankings

- Form the augmented coefficient matrix [A b] and solve using rref
-
- >> rref([A b]) gives us
-
- 1.0000 0 0 0 0.5040
- 0 1.0000 0 0 0.5278
- 0 0 1.0000 0 0.2183
- 0 0 0 1.0000 0.7500

- So Cinncinati has a rating of .5040
- Conn has a rating of .5278
- DePaul has a rating of .2183
- Georgetown has a rating of .7500
-
- So #1 Georgetown, #2 Uconn, #3 Cinn

#4 DePaul,

Expanding the model

- If we had the data for N teams then it is easy to define N dimensional vectors w and L (w is wins and L is losses)
-
- So **t = w+L**
-
- We will call the N dimensional vector [r1, r2,…,rN]’ = **R** the ratings vector.
-
- We will define an N dimensional column vector **ones** = [1,1,…,1]’

Matrices

- We can define a matrix G as follows
- $G(i,j)$ = # of times team i played team j when $i \neq j$
- 0 if $i = j$
- Define a matrix T by $T(i,j) = t_i$ for $i = j$
- $= 0$ for $i \neq j$
- Note that we can find G and T from a schedule of games.

Matrix Algebra

- we can write the equations (2I +T)*R = ones +1/2*(w-L) +G*R

- which gives us (2I+T)*R –G*R = (2I +T –G)*R = ones +.5*(w-L)
-
- So in our example A = 2I + T – G and **x** =R and **b** = ones+.5(w-L)

advantages

- We know the Identity matrix
- T is known from the vector t
- The entries of G can be found on the schedule
- We can see a pattern that is true for any number of teams that we obtained by matrix algebra and we can work with any size system just as easily as a small system like our example.
- We can ask and answer questions such as: “ is there always a unique solution to our equations” regardless of the number of teams involved.

Function ratings=colley(W,n)

- % W is n x n matrix with W(i,j) as the number of wins by team i over team j
- % so losses by team i are sum of col i of W
- % and wins by team i are the sum of col i of W’
- loss=sum(W’); win=sum(W);
- total=win+loss ;
- T=diag(total); G=W+W’;
- A=2*eye(n)+T-G; b= ones(n,1)+.5*(win-loss)
- E=rref([A b]) R=E(:,n+1)

Momentum

- Many fans believe a team that is “hot” at the end of the season is more likely to do well in the tournament than a team that has lost some games near the end of the season.
- It is easy to incorporate this idea into the model by counting games near the end of the season more heavily than games at the beginning of the season.

How I used it in the course

- Started with a small set of games from the Big East (4 teams , 8 games)in class (solved 4 x4)
- Then Big East conference schedule until Mar 2 (15X15)
- Class had to build their own matrices from spread sheet I put on SAKAI of schedule until Mar 2 and solve using Matlab. They should have a Matlab script or function which would take matrix and output ranks.
- Updates Mar 6,Mar 10 (start of tournament) and two after Big East tournament (one with tournament games given triple weight)
- Weighting and updates illustrate how matrix algebra makes things easy.(simply use linear combinations of the matrices as inputs for the ranking function.)

ESPN

- Students had to fill out one bracket using Colley models and one with Massey models and hand them in.
- Then they had to submit one bracket to our group at ESPN as well and give it to me with explanations for picks.(they could fill out their entry any way they wanted but they had to explain how they arrived at their selections)
- Bonus points available for brackets in
- top 10% + 25 , top 20% +20 top 30% + 15
- Top 40% +10 top 50% +5 (on top of 100pts)
- 4 in top 10%,4 in top 20, 2 in top 30, 5 in top 40, 4 in top 50, 4 < top 50

How can you use this for a Course

- Several different models can be used and you can involve different concepts
- I used the Colley model , lets you introduce matrix algebra in a meaningful way
- You can start with small sets then handle larger groups with matrix algebra
- Reverse idea Winning percentage has uncoupled equations you try to find a coupled system.

Resources

- ESPN brackets (CBS also has this)
- UDEMY course

March MATHness at udemy.com by Tim Chartier   This a set of short lectures and activities about different ways of ranking. You'll find them at:

http://www.udemy.com/march-mathness

- Signing up is free. There is software there too.

And

Rankings for Colley and Massey rating with weights through the UDEMY course (VERY IMPORTANT) This was how the class got rankings for the tournament

- For data and ratings see www.masseyratings.com

What worked well

- Using Big East to start the ideas(after Feb 1 they had no non conference games so we could use conference results)
- Using ESPN as the bracket deadline. Can’t submit after brackets lock.
- ESPN shows how the real world cares (over 8 million entries)
- Many students had ideas about how to improve models based on this particular problems. Momentum, strength of schedule, point differentials etc.
- Based on evaluations students thought this was a good addition to the course.
- Students could talk about this problem with non math majors. Seems related to the real world.
- If you feel creative you can pick a theme song (Work Hard,Play Hard by Wiz Khalifa) and mimic Sports Center and such

Problems and Improvements

- This was essentially added to the course at the last minute so I only counted it as 5% of the grade , probably should be at least 10%
- Some problems with organization and what was required and when.
- If you started this project at the beginning of the class you could have weekly results with matrix for each week that could be added and weighted as you wish to find ratings.
- Calender: Be aware of when the selection week occurs and make sure your class will be in session then.

Other Models

- Massey model uses point differentials for ratings. So if team i beats team j by 7 points we have the equation ri-rj=7 for that game. This system will probably not be consistent so solve by using least squares. This model can be weighted.
- You could weight road wins more than home court wins.

Models vs Seeds

- I submitted 5 brackets(all Louisville wins)
- Massey weight 1-2-4-8 95.5% 6 upsets
- Massey web site 87.5% 1 upset
- avg weighted Colley and weighted Massey
- 85% 4 upsets
- Bias(Big East Lville V ND,Georgetown V Marquette) 83.9% 14 upsets
- Colley weight 1-2-4-8 78% 11 upsets

THANK YOU

- hilbert@ithaca.edu