Cadena de Márkov y el baseball

El baseball es un deporte conformado por 9 jugadores en cada equipo y consta de completar circuitos para generar puntos. La forma de puntaje es, pasa un bateador y tiene 3 posibilidades, hacer un Home run, correr entre 3 diamantes (Triplete, doble o simple) o tener un out. Si el jugador hace un home run, implica que el y en su caso donde exista otro jugador del mismo equipo en la cancha podrán terminar el circuito y sumar un punto por jugador. El juego consiste de 9 entradas y cada entrada se termina cuando se consigue 3 out.

Nosotros nos vamos a basar en eventos básicos para simular cuantos puntos puede conseguir un equipo y eso será mediante una matriz de transición de 24 estados por entrada. Estos 24 estados serán de la siguiente forma:

- En realidad son 8 estados por out, ya sea 0,1 o 2 out, ya que el tercer out termina la entrada.
- Vamos a necesitar un estado absorbente, que representa el tercer out, siendo un total de 25.

Figure 1: State space of the Markov chain; states are labeled in (B,O) format.

En la imagen podemos observar los primeros 24 estados, fila 1 es cero out y así sucesivamente. La forma de (i,j) donde i representa las posiciones de los jugadores en la cancha y j la cantidad de out que han pasado en la entrada.

Pongamos un ejemplo. Si queremos pasar de (123,1) a (0,0) no es posible ya que los outs no desaparecen, pero si podríamos pasar de (123,1) a (0,1) si hacemos un Home run (No sería la única posibilidad).

Al ser 9 entradas al final vamos a tener 9*24 = 216 estados, más el estado absorbente, nos da un total de 217 estados por juego lo que hace una matriz de 217x217. Y si sabemos que cada equipo tiene 9 jugadores, estos nos haría una matriz o un tensor de 9x217x217.

Vamos a necesitar otra matriz de las carreras y vamos a suponer que ningún jugador se vuela la base, entonces pueden avanzar cada vez que alguien batea y no hay outs, de igual manera solo se puede sumar puntos cuando no hay out.

La matriz de jugadores la vamos hacer de la siguiente forma. En nuestro código usamos la función creada “crea_Matriz” y hace lo siguiente:

- Hace la suma total de las estadísticas del jugador
- Con ese total saca la probabilidad (caso favorable entre caso total)
- Creamos una pequeña matriz de 8x8 que será replicada en toda la diagonal de la matriz de 217x217
- La matriz de 8x8 son las probabilidades de cambiar de estado sin que ocurra un out
  Ejemplo:
  Matriz8[1,] = c(h, b+s, d, t, 0, 0, 0, 0) empezando en (0,0)
  1. Quedar en (0,0) es la probabilidad de un home run
  2. Quedar en (1,0) es la probabilidad de hacer un simple o que te regalen la primer, por eso se suma.
  3. Quedar en (0,2) es la probabilidad de hacer un doble
  4. Quedar en (0,3) es la probabilidad de hacer un triplete
  5. El resto no es posible ya que no hay más jugadores en la cancha.

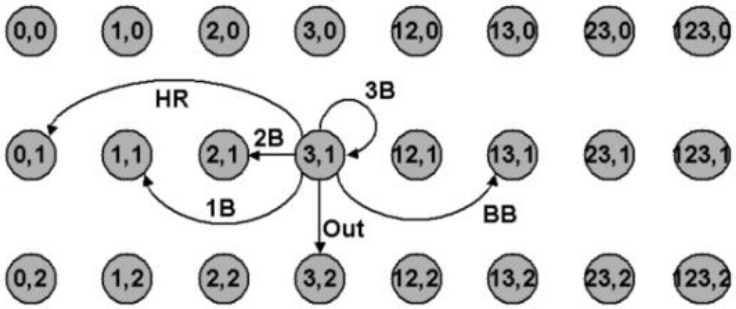

Figure 2: Potential transitions from state (3,1) using simplified baseball event model.

- Después Ajustamos la matriz según los out que han ocurrido y reemplazamos unos valores por la matriz de identidad
- Al final para el tercer out hacemos otros reemplazos con los valores de 1 para así crear el estado absorbente y el por último valor de la matriz igual se le da un valor de 1 ya que es el final del juego.

La matriz de Carrera se hace de una forma similar, excepto que la matriz de 8x8 usaremos la que venía con los datos. No hay necesidad de crearla.

Después para cada bateador, volvemos a calcular nuestro vector de probabilidad de situación basad en la matriz de probabilidad de transición de estados de jugadores . Si S es el vector de situación y P es la matriz, la fórmula es solo S: = S * P

También necesitamos hacer un seguimiento de cuántas carreras se anotaron en esos transiciones Si R es la matriz de carrea para cada transición, entonces R * P es el número esperado

de carreras que el jugador tendrá para cada transición y, S * (R * P) es el vector del número esperado de carreras que el jugador tendrá si batea en cada estado, ponderado por la probabilidad de cada estado. Entonces suma (S * (R * P)) da el número total de carreras esperadas cada que batea un jugador, si sumamos todas las sumas nos da el resultado del partido.

Los jugadores se eligen 9 de una base de forma aleatoria y en el orden que se toma es el orden en el que pasan los jugadores.

Conclusiones:

Basado en el paper, las simulaciones por Márkov pueden llegar a tener un margen de error de hasta 7%. Comprobando con nuestras simulaciones también nos damos cuenta que con la base usada, el orden de los jugadores no afectan tanto, como puede ser que tomes a los 9 jugadores como el mismo, todos por igual es con una data limitada a solo 9 jugadores. Con la data del mismo jugador puede variar más de 4 carreras, mientras con una data "PlayerData.csv" de más de 41 mil jugadores la diferencia es de menor de una carrea, solo es por fracción.

Los primeros 9 renglones del .csv son los 9 registros de la data del paper, y así enseñar las simulaciones con esos 9, misma que hacer el documento.

Hay formas de mejorar la simulación, por ejemplo hacer la data por equipo, para ser más realistas, meter los eventos secundarios, como robar base, etc.

Se podría hacer comparativa entre equipos y sería ideal poder meter a la defensa para simular un partido completo y no solo a la ofensiva.

Referencias

- Base de datos

http://www.seanlahman.com/baseball-archive/statistics/

- An Intuitive Markov Chain Lesson From Baseball

https://pubsonline.informs.org/doi/pdf/10.1287/ited.5.1.47

- FINDING BETTER BATTING ORDERS

http://www.pankin.com/markov/btn1191.htm

-The Markov Chain Model of Baseball

http://statshacker.com/blog/2018/05/07/the-markov-chain-model-of-baseball/