Métodos Numéricos
Segundo Cuatrimestre 2016
Trabajo Práctico 3

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Metnumball[1]

Contexto y motivación

Durante las últimas décadas el uso de estadísticas en los deportes se ha incrementado notablemente. Los indicadores muchas veces son utilizados para medir el rendimiento de los equipos en distintos aspectos así como también el desempeño individual de cada jugador. Dichos indicadores se usan con el fin de comparar pero también para saber qué aspectos del juego son fortalezas o debilidades con el fin de aprovecharlas o mejorar.

Uno de los deportes masivos pioneros en adoptar la medición de estadísticas con los objetivos mencionados fue el béisbol. En la liga de Estados Unidos, la MLB, la idea de usar programas estadísticos para analizar a los equipo de béisbol surge entre las décadas del 70 y del 80 acuñándose el término *sabermetrics* en el año 1980. Dicho término significa para su creador "the search for objective knowledge about baseball"; sin embargo, el uso de estadísticas para medir el desempeño de los equipos data de los años 40 y 50. Este tipo de mediciones resultan de gran utilidad en deportes con roles tan delimitados como el béisbol y el fútbol americano (donde las estadísticas comenzaron a usarse activamente entre los 90 y los 2000) pero la idea también fue trasladada al básquetbol. Si bien la NBA (liga de Estados Unidos) presenta una de las versiones del básquet con más juego de rol, es un deporte donde cada jugador cumple diversos roles, algunos ofensivos y otros defensivos, y donde la dinámica de equipo es un factor importante. Es por esto que durante la década del 90 y 2000 se comenzaron a diseñar métricas que midiesen el desempeño de los equipos y los jugadores, principalmente basándose en los indicadores que por décadas siempre se relevaron en la NBA como puntos promedio por partido, cantidad de asistencias o porcentaje de efectividad en triples entre muchas otras.

Actualmente todas las franquicias en la NBA entienden la importancia del uso de estadísticas para entender su juego y mejorar así como también determinar qué incorporaciones pueden resultar más útiles para el equipo dada la cantidad limitada de recursos que hay para contrataciones. En este trabajo buscaremos plantear métricas que permitan predecir el rendimiento de los equipos en temporada regular. Dicho rendimiento será dado a partir del win rate que se utiliza para rankear a los equipos y determinar quiénes clasifican a los tan ansiados play-offs.

El problema

La NBA es conocida por sus estadísticas y "datos curiosos" sobre el desempeño de los jugadores donde por ejemplo Rasheed Wallace ostenta el record de mayor cantidad de technical fouls (que pueden ser causadas por insultar, pelear o usar tabaco entre otras razones) en una temporada con 41 en 80 partidos durante la temporada 2000-01. Con el fin de predecir el rendimiento de un equipo utilizaremos las mediciones más comunes como cantidad de triples

[1] En alusión al término *moneyball* con el que también se conoce a sabermetrics y que da nombre al libro y película que hicieron famosa a la técnica.

o porcentaje de tiros libres convertidos entre otras. El objetivo será plantear dos métricas, una basada en estadísticas a nivel equipo y otra a nivel jugadores, fundamentarlas y analizar su rendimiento comparándolas con otras métricas ampliamente usadas. Las estadísticas serán tomadas del sitio basketball-reference[2] donde es posible ver gran cantidad de indicadores incluyendo algunos relativamente complejos de calcular como Player Efficiency Rate.

El objetivo será poder predecir la proporción de victorias sobre el total de partidos jugados, también conocida como *win rate*. Para esto se usará la técnica de Cuadrados Mínimos Lineales obteniendo funciones a partir de mediciones sobre el desempeño de los equipos y jugadores.

En el caso de la métrica a nivel equipo las estadísticas disponibles se encuentran detalladas en el Cuadro 1 las cuales consideran para cada equipo el total o promedio de los valores sobre todos los partidos de la temporada regular. En el sitio[3], en una tabla se encuentran los datos de cada equipo y en otra tabla las mismas estadísticas sumando o promediando los valores para los rivales sobre todos los partidos que jugó. Además, se presenta una métrica conocida como Four Factors [4] con el fin de tener un punto de comparación.

En cuanto a la métrica a nivel jugadores[5] las estadísticas se detallan en el Cuadro 2 donde los indicadores son totales o promedio según correspondan y además se indica el equipo para el que jugaron y obtuvieron esos resultados. Dado que se busca predecir los resultados del equipo, eventualmente, será necesario condensar los valores obtenidos por los jugadores de alguna manera como promediar o promediar pesadamente según la cantidad de minutos jugados. Esta idea es por ejemplo usada para el cálculo del PER (Player Efficiency Rate[6]) de manera de tener un valor promediado sobre todo el equipo por partido.

Tanto para Four Factors como para PER no se debe asumir que se trata de métricas perfectas sino que al tratar de capturar el rendimiento de un jugador o un equipo con un número (o un conjunto de ellos) se decide qué observar y qué importancia darle a cada aspecto tomando decisiones de alguna manera arbitrarias. En este sentido es que existen muchas críticas hacia estos indicadores como que por ejemplo que PER se concentra en aspectos ofensivos del juego pero deja de lado ciertas cuestiones defensivas sobre el desempeño de los jugadores. El objetivo de considerar estas métricas es usarlas como inspiración para plantear las propias y tener un punto de comparacón con el que contrastar las propuestas.

Para las dos métricas propuestas será necesario explicar qué criterios se usaron para formularlas. Algunas ideas para ganar conocimiento sobre el impacto de los datos que se tienen es analizar el nivel de correlación entre las estadísticas disponibles (por ejemplo midiendo covarianza entre las variables) y/o basarse en bibliografía del tema (que deberá ser citada debidamente). Tener en cuenta que gran parte de la información que se encuentra en internet no necesariamente es confiable. En caso de referenciar comentarios se espera que sean de alguien especializado en el tema como por ejemplo periodistas expertos en básquet. De todos modos se espera que el grupo sea crítico sobre la información que encuentre así como sobre las métricas propuestas fundamentando sus críticas y realizando experimentos con los datos que permitan validar o refutar sus conjeturas.

[2] http://www.basketball-reference.com/

[3] Para la temporada 2015/2016 es http://www.basketball-reference.com/leagues/NBA_2016.html

[4] http://www.basketball-reference.com/about/factors.html

[5] Para la temporada 2015/2016 es http://www.basketball-reference.com/leagues/NBA_2016_totals.html

[6] http://www.basketball-reference.com/about/per.html

Técnicas a utilizar y métricas de evaluación

La técnica de Métodos Numéricos a utilizar para proponer los modelos es Regresiones Lineales/Cuadrados Mínimos Lineales (CML). Para determinar el modelo, se tiene una serie de N observaciones $(x_{(i)}, y_{(i)})$, con $x_{(i)} \in \mathbb{R}^k$ el vector de *features* e $y_{(i)} \in \mathbb{R}$ la variable dependiente (el win rate en nuestro caso). Luego, el modelo consiste en encontrar los parámetros (lineales) que definen $y_{(i)} = f(x_{(i)}) + \epsilon_i$, $i = 1, \ldots, N$, donde ϵ_i es el error de la medición i-ésima, y que minimizan el error de la aproximación en el sentido de CML.

Dado un conjunto de datos $\{(x_{(i)}, y_{(i)}\}_{i=1,\ldots,N}$ será necesario considerar distintas hipótesis sobre la función f (por ejemplo, considerar polinomios de distinto grado) que dan lugar a distintos modelos. Para poder decidir entre los mismos, es necesario considerar alguna métrica de evaluación. Se sugiere considerar el *Mean Squared Error* (MSE)[7]. Dado un modelo $\hat{f}$ de f y una observación $(x_{(i)}, y_{(i)})$, se define $\hat{y}_{(i)} = \hat{f}(x_{(i)})$ y $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$. Con estas definiciones, se puede calcular el MSE del modelo $\hat{f}$ como

$$MSE(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} e_{(i)}^2.$$

Otra posibilidad consiste en considerar $\max\limits_{i \in \{1 \ldots N\}} e_{(i)}^2$ de manera que se busque minimizar el máximo error. Ya sea que se elija uno de estos criterios u otro propuesto por el grupo será necesario justificar brevemente la elección.

Esta metodología sirve para evaluar cuán bien ajusta el modelo en función de los datos de entrenamiento utilizados. Sin embargo, en un contexto de modelos predictivos se corre el riesgo de caer en el conocido *overfitting*. Para evitar este fenómeno, se puede considerar la técnica de *cross-validation* (CV). Es decir, particionar el conjunto de datos y variar la composición de la base de entrenamiento (*training*) y las observaciones consideradas como *test*. Una vez obtenido el modelo $\hat{f}$, se toman las observaciones en el conjunto de test, se aplica el modelo y se evalúa la métrica de evaluación obtenida. La métrica final para el modelo $\hat{f}$ consiste en tomar alguna medida sobre los resultados obtenidos para cada combinación de training/test considerado.

Es importante notar que si se consideran datos de diferentes temporadas para train/test se debe tener en cuenta que el conjunto sea representativo. Es sabido que las dinámicas del juego cambian a lo largo del tiempo en los deportes y la NBA no es la excepción de modo que si uno entrena con datos de un cierto período de años y evalúa en otro muy posterior es posible que estos efectos hagan que los resultados no sean demasiado confiables. Será necesario explicar datos de qué períodos se consideran y justificar brevemente la elección.

Enunciado

El Trabajo Práctico consiste como punto de partida considerar los datos disponibles en basketball-reference[8] y formular dos métricas para predecir la proporción de victorias sobre el total de partidos. Para ello, se deberá utilizar CML como técnica de análisis y modelado, tanto a nivel descriptivo de los datos como a nivel predictivo de resultados futuros. Para el

[7] Notar que MSE es dependiente de la escala.

[8] http://www.basketball-reference.com/

desarrollo de los métodos se podrá considerar como posibles lenguajes MATLAB, Python y/o C++. Se remarca que, a diferencia de trabajos anteriores, no es necesario realizar toda la implementación desde cero y es posible utilizar rutinas provistas por dichos lenguajes. El objetivo principal de este trabajo se centra en la aplicación de las técnicas de CML a una temática práctica concreta y en la correspondiente experimentación necesaria para evaluar los desarrollos. Se deben reportar dos métricas (una basada en estadísticas de equipos y otra a nivel de jugadores) de forma completa y comparar los resultados con los obtenidos al usar Four Factors y PER respectivamente.

Junto con el enunciado se proveen una serie de archivos con información sobre distintas temporadas tanto para equipos como para jugadores y scripts en bash para obtener valores de PER por equipo y un indicador condensado de Four Factors. Además, se presentan dos scripts para extraer determinados indicadores tanto a nivel equipo como jugador en un cierto período de tiempo (indicado en un archivo de parámetros). Se brindan además las tablas para las últimas 30 temporadas y los scripts permiten extraer datos de dichas tablas. No debe asumirse que esos son todos los datos disponibles, pueden sin problemas tomarse datos de temporadas anteriores. Una vez aplicados dichos scripts, los resultados podrán eventualmente resultar más fáciles de utilizar en el código que realice el modelo.

Si es necesario el grupo puede plantear sus propias herramientas de scripting para filtrar los datos de acuerdo a sus necesidades. En este sentido, es posible que los grupos compartan, a través de la lista de alumnos de la materia, herramientas de preprocesamiento y extracción de datos con otros grupos. Es importante evitar que las herramientas compartidas contengan información particular de las métricas planteadas y la experimentación a realizar.

Los resultados deben ser volcados en un informe con la estructura habitual. Sin embargo, en este caso es obligatorio escribirlo utilizando el template de la revista *Electronic Notes on Discrete Mathematics* (ENDM). Además, el informe no podrá exceder las 10 páginas de longitud y, por lo tanto, los resultados tienen que ser presentados y condensados de forma adecuada. Notar que esto no significa que la experimentación debe ser acotada, sino todo lo contrario: es importante realizar muchos experimentos y mostrar los que resulten representativos. Como en los demás trabajos, es importante proveer la información necesaria para poder replicar todos los experimentos, ya sea que se encuentren en el informe o no.

Por último, este trabajo tendrá una presentación oral frente a un grupo de docentes que será evaluada como una parte adicional de la nota. Para la misma, cada grupo diseñará una presentación incluyendo los desarrollos y resultados que considere interesantes, plasmados en el informe, y dispondrá de 15 minutos para exponerlo. La exposición puede ser de la totalidad o de un subconjunto de los integrantes, y esta decisión queda a elección del grupo. Una vez finalizada la misma, se llevará a cabo un coloquio donde los integrantes del grupo responderán a las preguntas realizadas. Cabe mencionar que los docentes podrán elegir qué alumno debe responder, con lo cual es importante que todos los integrantes estén al tanto de todas las decisiones tomadas.

Fechas de entrega

- *Formato Electrónico:* Miércoles 16/11, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP3]` seguido de la lista de apellidos de los integrantes del grupo.

- *Confirmación presentación oral:* Viernes 18/11, por correo electrónico.
- *Presentación oral:* Lunes 21/11, en horario a determinar luego de la confirmación. Será en horario de clase de la materia.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.

Cuadro 1: Indicadores sobre el desempeño de los equipos.

MP	Minutes Played
FG	Field Goals
FGA	Field Goals Attempts
FG %	Field Goal Percentage
3P	3-Point Field Goals
3PA	3-Point Field Goals Attempts
3P %	3-Point Field Goals Percentage
2P	2-Point Field Goals
2PA	2-Point Field Goals Attempts
2P %	2-Point Field Goals Percentage
FT	Free Throws
FTA	Free Throw Attempts
FT %	Free Throw Percentage
ORB	Offensive Rebounds
DRB	Deffensive Rebounds
TRB	Total Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points
PTS/G	Points Per Game

Cuadro 2: Indicadores sobre el desempeño de los jugadores.

POS	Position
Age	Age during that season
G	Games Played
GS	Games Started
MP	Minutes Played
FG	Field Goals
FGA	Field Goals Attempts
FG %	Field Goal Percentage
3P	3-Point Field Goals
3PA	3-Point Field Goals Attempts
3P %	3-Point Field Goals Percentage
2P	2-Point Field Goals
2PA	2-Point Field Goals Attempts
2P %	2-Point Field Goals Percentage
eFG %	Effective Field Goal Percentage
FT	Free Throws
FTA	Free Throw Attempts
FT %	Free Throw Percentage
ORB	Offensive Rebounds
DRB	Deffensive Rebounds
TRB	Total Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points