Métodos Numéricos
Primer Cuatrimestre 2016
Trabajo Práctico 1

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

(No) Todo Pasa

Contexto y motivación

Las competencias deportivas, en todas sus variantes y disciplinas, requieren casi inevitablemente la comparación entre competidores mediante la confección de *Tablas de Posiciones* y *Rankings* en base a resultados obtenidos en un período de tiempo determinado. Estos ordenamientos de equipos están generalmente (aunque no siempre) basados en reglas relativamente claras y simples, como proporción de victorias sobre partidos jugados o el clásico sistema de puntajes por partidos ganados, empatados y perdidos. Sin embargo, estos métodos simples y conocidos por todos muchas veces no logran capturar la complejidad de la competencia y la comparación. Esto es particularmente evidente en ligas donde, por ejemplo, todos los equipos no juegan la misma cantidad de veces entre sí.

A modo de ejemplo, la NBA y NFL representan dos ligas con fixtures de temporadas regulares con estas características. Recientemente, el Torneo de Primera División de AFA se suma a este tipo de competencias, ya que la incorporación de la *Fecha de Clásicos* parece ser una interesante idea comercial, pero no tanto desde el punto de vista deportivo ya que cada equipo juega contra su *clásico* más veces que el resto. Como contraparte, éstos rankings son utilizados muchas veces como criterio de decisión, como por ejemplo para determinar la participación en alguna competencia de nivel internacional. En el caso de competencias en los estados unidos, las posiciones finales deterinan cuál es la prioridad entre los equipos para la elección de los nuevos jugadores que ingresan a la liga mediante el conocido proceso de *Draft.*Luego, la confección de los rankings finales de los equipos constituye un elemento sensible, afectando intereses deportivos y económicos de gran relevancia.

En un contexto de extremada desconfianza respecto a los manejos a nivel local, regional e internacional de las confederaciones de fútbol, en este trabajo nos proponemos estudiar el

comportamiento de otras métricas para la generación de rankings en competencias deportivas con el fin de brindar mayor transparencia y nivelar la competitividad, en un futuro, de nuestras ligas locales.

El problema

Existen en la literatura distintos enfoques para abordar el problema de determinar el *ranking* de equipos de una competencia en base a los resultados de un conjunto de partidos. En Govan et al. [5] se hace una breve reseña de varios de ellos, y e incluso los autores proponen uno nuevo.[1] Entre los métodos presentados se encuentra el denomiando *Colley Matrix Method* (CMM) [1, 5]. El método se basa en la *Regla de Laplace de sucesos* y solo requiere conocer el historial de partidos y los respectivos resultados (básicamente, quién ganó) de los mismos. Esta regla permite aproximar las probabilidades de eventos *booleanos*, en nuestro caso que un equipo gane o pierda un partido. En particular, si sobre k eventos observamos s casos exitosos, la regla establece que $(s+1)/(k+2)$ es un mejor estimador que el porcentaje estándar, s/k. En base a esta idea, el problema se reformula como la resolución de un sistema de ecuaciones lineales, que permite obtener estos estimadores y, por lo tanto, el ranking deseado.

Extendiendo la notación introducida en Govan et al. [5], sea $\Gamma = \{1, 2, \ldots, T\}$ el conjunto de participantes de la competencia. Para cada equipo $i \in \Gamma$ llamamos n_i all número total de partidos jugados por el equipo i, w_i al número de partidos ganados por el equipo i y, análogamente, l_i al número de partidos perdidos por el equipo i. Definimos también dados $i, j \in \Gamma$, $i \neq j$, n_{ij} al número de enfrentamientos entre i y j. Es importante destacar que el modelo asume que el empate no es un resultado posible.

El método CMM propone construir una matriz $C \in \mathbb{R}^{T \times T}$ y un vector $b \in \mathbb{R}^T$, tal que el ranking buscado $r \in \mathbb{R}^T$ es la solución del sistema $Cr = b$. Para el armado del sistema, se define

$$C_{ij} = \begin{cases} -n_{ij} & \text{si } i \neq j, \\ 2 + n_i & \text{si } i = j. \end{cases}$$

y $b_i = 1 + (w_i - l_i)/2$, $i \in \Gamma$.

Los detalles respecto a la formulación del sistema pueden ser consultados en Colley [1, Method]. Este método puede ser aplicado a una gran variedad de deportes y tipos de competencias, incluyendo información de conferencias, divisiones, etc. El objetivo central de este trabajo práctico consiste en estudiar el comportamiento del mismo, en conjunto con el análisis de algunos de los métodos que pueden ser utilizados para su resolución.

Como punto de comparación, se considerará (al menos) un método alternativo para generar rankings. Una opción es considerar el *porcentaje de victorias* (WP), donde el puntaje asignado al equipo $i \in \Gamma$ está dado por $w_i/(w_i+l_i)$. En caso de ser factible, es posible también incorporar el método que se aplique en la competencia elegida.

Enunciado

Se debe implementar un programa en `C` o `C++` que tome como entrada el detalle de los partidos de la competencia y calcule el ranking en función de los métodos mencionados en la sección

[1] Remarcamos que este no es el método involucrado en el TP1. Será visto en el segundo tercio de la materia.

anterior (CMM, WP ó el método elegido por el grupo). El formato de los archivos se detalla en la siguiente sección.

La matriz resultante del sistema planteado por el método CMM es *Simétrica y Definida Positiva* (ver, e.g., [1, Sección Method]) y, por lo tanto, es posible encontrar la Factorización de Cholesky para resolver el sistema. Luego, como parte obligatoria en relación a los métodos de resolución de sistemas de ecuaciones lineales se pide implementar:

- el método de Eliminación Gaussiana clásico (EG), y
- el método de Cholesky (CL).

Es importante incluir en el informe del TP, en la sección desarrollo, aquellas decisiones tomadas en función de la estructuras de datos utilizadas y las alternativas consideradas y descartadas durante el proceso. Además, sabemos que existen casos donde el algoritmo EG no puede encontrar una solución. Se pide incluir en el desarrollo una justificación sobre por qué el algoritmo funciona correctamente en el caso del método CMM.

La experimentación será divida en dos partes, cada una con sus respectivos ejes. En primer lugar, buscamos hacer una evaluación cuantitativa de los métodos de resolución de sistemas lineales considerados, i.e., EG y CL, en términos del tiempo de cómputo y el tamaño de los sistemas a resolver. En particular, se pide comparar, para distintos tamaños de matrices, el tiempo de cómputo requerido para cada método en el contexto donde la información de la matriz del sistema (C) se mantiene invariante, pero varía el término independiente (b). Si las instancias obtenidas de datos reales no permiten notar diferencias significativas, se puede reformular el experimento utilizando instancias artificiales generadas convenientemente. Justificar cómo se generan estos datos y por qué es posible tomar esta decisión para este aspecto del análisis.

La segunda parte de la experimentación se centra en el análisis cualitativo respecto del comportamiento de los métodos CMM, WP o el elegido por los integrantes del grupo. Entre los experimentos a realizar, se pide como mínimo analizar los siguientes aspectos e intentar responder las siguientes preguntas:

- Utilizar principalmente datos de competencias reales que permitan identificar características distintivas de los métodos, y relacionarlas con eventos que ocurren en los mismos. Comparar los rankings obtenidos por cada uno de los métodos considerados.
- El método CMM es *justo*? Es decir, es posible que el resultado de un partido entre dos equipos afecte indirectamente el ranking de un tercero?
- Dados los resultados de todos los partidos considerados en la competencia, y un equipo particular. Determinar una estrategia que permita obtener la mayor posición posible, buscando minimizar el número de partidos ganados.[2]

En todos los casos es obligatorio fundamentar los experimentos planteados, proveer los archivos e información necesarios para replicarlos, presentar los resultados de forma conveniente

[2]No es necesario que la cantidad de partidos ganados sea la mínima, pero sí que la estrategia planteada trate de minimizar este aspecto.

y clara y analizar los mismos con el nivel de detalle apropiado. En caso de ser necesario, es posible también generar instancias artificiales con el fin de ejemplificar y mostrar un comportamiento determinado.

En la era del auge de *Big Data*, el presente trabajo puede ser enmarcado dentro de un área en continuo crecimiento y desarrollo denominada *Sports Analytics*, enfocada en aplicar técnicas estadísticas, de inteligencia articial y optimización a los deportes. En este contexto, se pide que el grupo vea las películas *Moneyball* y *Trouble with the curve*, para luego analizar y reflexionar sobre la siguiente afirmación:

La utilización de técnicas avanzadas de análisis de datos son imprescindibles para mejorar cualquier deporte.

Como puntos opcionales para incluir en el desarrollo y/o experimentación, se consdiera lo siguiente:

1. Proponer y discutir (al menos) una forma alternativa de modelar el empate entre equipos en CMM.

Parámetros y formato de archivos

El programa deberá tomar por línea de comandos tres parámetros. El primero de ellos contendrá el path al archivo de entrada con los partidos y resultados de la competencia; el segundo la salida con el ranking correspondiente, y el tercero indicando el método a considerar (0 CMM-EG, 1 CMM-CL, 2 WP).

El archivo de entrada contiene primero una línea con información sobre la cantidad de equipos (n), y la cantidad de partidos totales a considerar (k). Luego, siguen k liñeas donde cada una de ellas representa un partido y contiene la siguiente información: identificador de fecha (es un dato opcional al problema, pero que puede ayudar a la hora de experimentar, un `string`), equipo i, goles equipo i, equipo j, goles equipo j.

A continuación se muestra el archivo de entrada con la información del ejemplo utilizado en Govan et al. [5]:

```
6 10
1 1 16 4 13
1 2 38 5 17
1 2 28 6 23
1 3 34 1 21
1 3 23 4 10
1 4 31 1 6
1 5 33 6 25
1 5 38 4 23
1 6 27 2 6
1 6 20 5 12
```

Es importante destacar que, en este último caso, los equipos son identificados mediante un número. Opcionalmente podrá considerarse un archivo que contenga, para cada equipo, cuál es el código con el que se lo identifica.

Una vez ejecutado el algoritmo, el programa deberá generar un archivo de salida que contenga una línea por cada equipo (n líneas en total), acompañada del puntaje obtenido por el algoritmo CMM/WP/método alternativo.

Para instancias correspondientes a resultados entre equipos, la cátedra provee algunas opciones con información real de resultados en distintas competencias. Desde ya que cada grupo puede buscar/generar sus propios conjuntos de datos en caso que así lo considere. En [3] se provee un extenso set de datos con resultados históricos de la liga ATP de tenis profesional, divididos por año. Si bien los archivos contienen estadśiticas detalladas sobre los partidos del circuito, en nuestro caso solo se necesitan un subconjunto muy reducido de los mismos. Por otro lado, en [4] se proveen resultados detallados para disintas ligas, profesionales y universitarias, de los Estados Unidos. Si bien es facil interpretar los archivos, la cátedra provee junto con este enunciado sripts en python para poder traducir los archivos obtenidos en cada uno de estos repositorios al formato requerido por el TP[3]. Finalmente, otra alternativa es considerar el repositorio DataHub [2], que contiene información estadística y resultados para distintas ligas y deportes de todo el mundo. En estes caso, no se proveen herramientas adicionales para su pre-procesamiento.

Junto con el presente enunciado, se adjunta una serie de scripts hechos en `python` y un conjunto instancias de test que deberán ser utilizados para la compilación y un testeo básico de la implementación. Se recomienda leer el archivo `README.txt` con el detalle sobre su utilización.

Fechas de entrega

- *Formato Electrónico:* Jueves 7 de Abril de 2016, 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 `[TP1]` seguido de la lista de apellidos de los integrantes del grupo.
- *Formato físico:* Viernes 8 de Abril de 2016, a las 18 hs. en la clase de laboratorio.

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

Referencias

[1] Colley rankings. `http://colleyrankings.com`.

[2] Datahub. `http://datahub.io`.

[3] Jeff sackmann atp tennis rankings. `http://github.com/JeffSackmann/tennis_atp`.

[4] Massey ratings. `http://masseyratings.com/data.php`.

[5] Angela Y. Govan, Carl D. Meyer, and Rusell Albright. Generalizing google's pagerank to rank national football league teams. In *Proceedings of SAS Global Forum 2008*, 2008.

[3] Los mismos es opcional. En caso de encontrar algun error/bug en los mismos, por favor comunicarlo a la brevedad a la lista de docentes de la materia.