Métodos Numéricos
2do Cuatrimestre 2021
Trabajo Práctico 1

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

Sistemas de ecuaciones lineales

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 (*National Basketball Association*) y la NFL (*National Football League*) representan dos ligas con fixtures de temporadas regulares con estas características, como así también el ranking de la ATP (*Association of Tennis Professionals*) posee el suyo.

En los últimos tiempos, hasta el Torneo de Primera División de AFA (Asociación del Fútbol Argentino) ha tenido versiones en donde se pareció a este tipo de competencias en donde los fixtures para los equipos no eran equivalentes.

Como contraparte, estos rankings son utilizados muchas veces como criterio de decisión, por ejemplo para determinar la participación en alguna competencia de nivel internacional. En el caso de competencias en los Estados Unidos, las posiciones finales determinan 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, e incluso los autores proponen uno nuevo.[1] Entre los métodos presentados se encuentra el denominado *Colley Matrix Method* (CMM) [1, 5]. El método se basa en la *Regla de Laplace de sucesos*

[1] Remarcamos que este no es el método involucrado en el este TP y será visto más adelante en la materia.

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 $\frac{s+1}{k+2}$ es un mejor estimador que el porcentaje estándar $\frac{s}{k}$. 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 equipos/participantes de la competencia. Para cada equipo $i \in \Gamma$ llamamos n_i al 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 y por lo tanto $n_i = w_i + l_i$

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 $r \in \mathbb{R}^T$ sea la solución del sistema $C\,r = b$. Para el armado del sistema, se define $\forall i, j \in \Gamma$:

$$C_{ij} = \begin{cases} -n_{ij} & \text{si } i \neq j, \\ 2 + n_i & \text{si } i = j. \end{cases} \tag{1}$$

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

Los detalles respecto a la formulación del sistema pueden ser consultados en Colley [1]. 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 **deberá** considerar (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 $\frac{w_i}{w_i + l_i}$. 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++` 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 anterior (CMM, WP y el método elegido por el grupo). El formato de los archivos de entrada/salida se detalla en la siguiente sección.

Como parte **obligatoria** en relación a los métodos de resolución de sistemas de ecuaciones lineales, se pide implementar el método de Eliminación Gaussiana (EG) sin permutaciones.

Previamente, **se deberán** estudiar las características de la matriz involucrada en CMM y responder a lo siguiente:

1. ¿Qué tipo de matriz resulta C?
2. ¿Cómo se garantiza la aplicabilidad de EG sin intercambio de filas o columnas?
3. ¿Qué se puede decir sobre la estabilidad de los cálculos?

Además, sabemos que existen casos donde el algoritmo EG no puede encontrar una solución. Se **debe** incluir en el desarrollo una justificación sobre por qué el algoritmo funciona correctamente en el caso del método CMM.

También, **se pide** realizar un informe utilizando como guía las pautas de laboratorio de la materia conteniendo la experimentación pedida en la siguiente sección. Es importante incluir en la sección desarrollo del informe del trabajo práctico, las alternativas consideradas y descartadas para cada uno de los métodos utilizados.

Se **recomienda** fuertemente en todos los casos comparar los resultados intermedios utilizando python, Matlab/Octave o R.

Experimentación

Se deberá realizar tanto un análisis **cualitativo** como **cuantitativo** de los métodos vistos en el trabajo práctico.

Análisis cuantitativo

Se **pide**, analizar y reportar los errores obtenidos en el ranking analizando el error absoluto con respecto a los test de la cátedra y los propios generados por el grupo en donde se evidencien problemas.

Análisis cualitativo

Analizar el comportamiento de los métodos CMM, WP y el elegido por los integrantes del grupo. Entre los experimentos a realizar, se **debe** como mínimo analizar los siguientes aspectos e intentar responder las siguientes preguntas:

1. ¿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?
   ¿Cómo afecta la dificultad del *schedule* al ranking de los equipos y que consecuencias tiene esto en la calidad de los resultados? Para las preguntas anteriores, se sugiere diseñar casos de torneos artificiales que faciliten el estudio inicial de los métodos en sí.

2. Utilizar 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.

3. 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 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.

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

Puntos opcionales

Como puntos opcionales para incluir en el desarrollo y/o experimentación sobre el método CMM, se considera lo siguiente:

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

2. La matriz resultante del sistema planteado por el método CMM tiene una característica que permite encontrar una factorización para resolver el sistema de forma más eficiente. Implementar y comprar con el método de Eliminación Gaussiana.

3. Implementar una estructura de matrices ralas que sea eficiente en espacio y en tiempo para la tarea que se busca realizar sobre la matriz de CMM. ¿En qué situaciones se justifica esta implementación?

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, 1 WP, 2 alternativo. Ejemplo:

```
$ ./tp1 partidos.csv  ranking.out 0
```

El archivo de entrada contiene primero una línea con información sobre la cantidad de participantes/equipos T, y la cantidad de partidos totales a considerar P. Luego, siguen P líneas donde cada una de ellas representa un partido y contiene la siguiente información separada por espacios o tabulador: identificador de fecha (es un dato opcional al problema de tipo `int` o `string`, pero que puede ayudar a la hora de experimentar), 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. Ejemplo:

```
0.359707
0.615978
0.668661
0.314936
0.501544
0.539174
```

Además, se **deberá** entregar un archivo `README` conteniendo la instrucciones de compilación y ejecución, y también ejemplos de invocación del programa.

Conjuntos de datos

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.

1. 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 estadísticas detalladas sobre los partidos del circuito, en nuestro caso solo se necesitan un subconjunto muy reducido de los mismos.

2. Por otro lado, en [4] se proveen resultados detallados para distintas ligas, profesionales y universitarias, de los Estados Unidos. Si bien es fácil interpretar los archivos, la cátedra provee junto con este enunciado *scripts* en `python` para poder traducir los archivos obtenidos en cada uno de estos repositorios al formato requerido por el TP[3].

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 este caso, no se proveen herramientas adicionales para su preprocesamiento.

Casos de test

La cátedra proveerá un conjunto de tests con archivos de entrada y salida esperada. Para aprobar el trabajo, los mismos **deberán** funcionar correctamente en sus implementaciones con una tolerancia máxima de 10^{-4} medida en error absoluto con respecto a los valores de ranking proporcionados.

Fechas de entrega

- *Formato Electrónico:* Jueves 9 de Septiembre hasta las 23.59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`.

    - El asunto del correo debe comenzar con el texto `[TP1]` seguido de la lista de apellidos de los integrantes del grupo separados por punto y coma `;`.
    - Se ruega no sobrepasar el máximo permitido de archivos adjuntos de 20MB. Tener en cuenta al realizar la entrega de no ajuntar bases de datos disponibles en la web, resultados duplicados o archivos de backup.

[3] Los mismos son opcionales. En caso de encontrar algún error en los mismos, por favor comunicarlo a la brevedad a la lista de docentes de la materia.

- *Recuperatorio:* jueves 30 de Septiembre hasta las 23.59 hs, enviando el trabajo corregido a `metnum.lab@gmail.com`.
- Pautas de laboratorio:
  `https://campus.exactas.uba.ar/pluginfile.php/163805/course/section/22657/pautas.pdf`

Importante: El horario es estricto. No se considerarán correos enviados después de hora.

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`.

[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.