Máster en Data Science and Big Data
Inteligencia Colectiva y
Sistemas de Recomendación
Curso 18-19

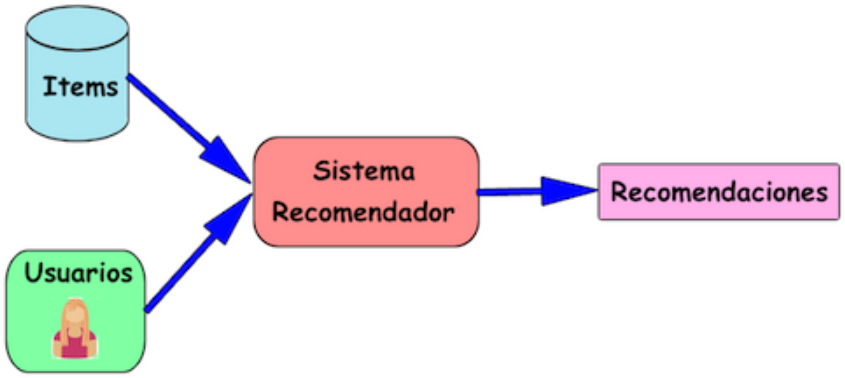
Javier Fernández Rodríguez

Abril 2019

Índice

1 Sistemas de recomendación

Un sistema de recomendación es un sistema inteligente que proporciona a los usuarios una serie de sugerencias personalizadas (recomendaciones) sobre un determinado tipo de elementos (items). Los sistemas de recomendación estudian las características de cada usuario y mediante un procesamiento de los datos, encuentran un subconjunto de items que pueden resultar de interés para el usuario.



Los sistemas de recomendación se pueden clasificar en 4 tipos: filtrado basado en contenido, filtrado demográfico, filtrado colaborativo y filtrado híbrido. En nuestro caso nos centraremos en el **filtrado colaborativo** que consiste en ver qué usuarios son similares al usuario al que hay que realizarle las recomendaciones y a continuación, recomendar aquellos items que (aún) no han sido votados por el usuario dado y que han resultado bien valorados por los usuarios similares. Además, nos interesaremos también en encontrar los items similares a uno concreto.

2 Identificación del problema

Como hemos dicho, la finalidad de un sistema de recomendación es predecir la valoración que un usuario va a hacer de un ítem que todavía no ha evaluado. Tras la predicción de las valoraciones faltantes, nuestro objetivo es ofrecer al usurario una lista de items ordenados por la puntuación obtenida. En el mundo del fútbol es habitual hablar de estadísticas y valoraciones, luego usaremos dicho campo para realizar nuestro análisis.

Hoy en día, muchos periodistas deportivos llenan páginas de periódicos y horas de radio hablando de los partidos que se han jugado el fin de semana y valorando la actuación de los diferentes jugadores. Actualmente, muchas aplicaciones y simuladores de fútbol dependen de las valoraciones que proporcionan los especialistas en este campo a los diferentes futbolistas, es el caso de **Biwenger, el mánager de fútbol fantasy oficial de AS y Cadena Ser** en el cual múltiples usuarios crean su propio equipo con jugadores de los diferentes equipos de la Liga Española de fútbol y, en función de las puntuaciones que periodistas del periódico AS y la Cadena Ser otorgan a los diferentes jugadores, alcanzan un mayor puntaje en una jornada determinada. Jornada tras jornada, los usuarios compiten entre sí de manera que el usuario ganador es aquel cuyo equipo ha recibido unas mejores valoraciones por los especialistas/periodistas de las fuentes anteriores a lo largo de toda la temporada.

2.1 Conjunto de datos

Pues bien, nuestro conjunto de datos va a ser el siguiente:

Se trata de un conjunto de datos con filas y columnas donde cada fila corresponde a las puntuaciones que ha proporciona un determinado periodista a diferentes jugadores de la Liga de Fútbol Profesional (*matriz de puntuaciones*). Así pues, tendremos un set de datos 6×6 donde el nombre de las filas hace referencia al nombre de los distintos periodistas y el nombre de las columnas hace referencia a los futbolistas de la Liga Española que han sido valorados. Las puntuaciones que recibe cada jugador oscilan entre el 1 y el 5, siendo 5 la máxima valoración posible.

Ahora bien, en nuestro conjunto de datos aparecerán columnas con registros faltanes (NaN), se debe a que dicho periodista no ha realizado una valoración sobre el futbolista en cuestión; se convertiran en los registros a predecir.

Futbolistas / Periodistas	Banega	Lo Celso	Messi	Modric	Ramos	Suarez
Cristobal Soria	3.0	NaN	NaN	3.0	3.5	4.5
Julio Maldonado	NaN	4.5	5.0	3.5	3.0	3.5
Lluis Flaquer	3.5	3.5	5.0	1.0	2.0	5.0
Manolo Lama	3.5	NaN	3.5	4.5	5.0	3.5
Manu Carreño	3.0	4.0	3.0	2.0	3.5	3.0
Tomas Roncero	2.5	4.0	4.0	5.0	NaN	3.5

2.2 Objetivos. Problemas a estudiar

Trataremos de abordar las siguientes situaciones:

- **Determinar la similitud entre futbolistas.** Los distintos jugadores han recibido una puntuación de acuerdo a su actuación el pasado fin de semana, luego parece interesante estudiar qué jugadores son similares a uno dado de acuerdo a las puntuaciones recibidas por los distintos periodistas.
- **Filtrado colaborativo: predecir nuevas valoraciones.** Como hemos comentado, hay periodistas que no han realizado valoraciones a algunos jugadores, luego nuestra tarea será predecir dicha valoración. Para llevar a cabo esta predicción, tendremos que calcular la similitud entre usuarios (periodistas en nuestro caso) que no será mas que un valor que nos diga el grado de similitud que tenemos con otro/s usuarios. Esta similitud la calcularemos con alguna de las siguientes **métricas**: coeficiente de Pearson, diferencia cuadrática media, distancia euclídea modificada, ...

- **Realizar las recomendaciones.** Una vez que tenemos las valoraciones que cada periodista ha realizado a cada jugador, recomendaremos a cada uno de ellos un listado de jugadores ordenados por la valoración obtenida.

3 Similitud entre items

Una manera de calcular la similitud entre futbolistas es calcular la correlación entre ellos en función de las valoraciones que les dan los periodistas (usuarios). Una forma de hacerlo en *python* es usando la función ***np.corrcoef***, que calcula el coeficiente de Pearson entre cada par de items. Dicho coeficiente toma valores entre el intervalo [-1,1] y mide la relación entre un par de variables. La matriz resultante es una matriz de tamaño m × m, donde el elemento M_{ij} representa la correlación entre el item i y el item j.

En la implementación en *python*, no podemos usar la función dicha pues nuestra tabla presenta registros faltantes. Una forma de subsanar este problema es recurrir a la función ***corr()***. Esta función nos devuelve la matriz de correlación del conjunto de datos dado, sin embargo, convertiremos dicha matriz a una del tipo *numpy.ndarray* pues nos interesrará tener dicho resultado en tal formato, como veremos a continuación, para posteriores análisis.

```
corr_matrix = np.array(data.corr())
```

```
corr_matrix
```

```
array([[ 1.        , -0.8660254 ,  0.2548236 , -0.46428571,  0.        ,
         0.40006613],
       [-0.8660254 ,  1.        ,  0.        ,  0.58321184,  0.65465367,
        -0.70710678],
       [ 0.2548236 ,  0.        ,  1.        , -0.22450166, -0.69526879,
         0.70034929],
       [-0.46428571,  0.58321184, -0.22450166,  1.        ,  0.83653809,
        -0.5       ],
       [ 0.        ,  0.65465367, -0.69526879,  0.83653809,  1.        ,
        -0.57547048],
       [ 0.40006613, -0.70710678,  0.70034929, -0.5       , -0.57547048,
         1.        ]])
```

Una vez tenemos la matriz, si queremos encontrar futbolistas similares a uno concreto, solo tenemos que encontrar los futbolistas con una correlación alta con el jugador dado. Consideraremos pues el futbolista **Leo Messi** y fijaremos nuestro objetivo en conocer la similitud entre este futbolista y el resto.

```
favoured_futbolista = 'Messi'
favoured_futbolista_index = list(futbolista_index).index(favoured_futbolista)
P = corr_matrix[favoured_futbolista_index]
```

```
a = pd.Series(P, index=futbolista_index)
a
```

```
Futbolistas
Banega      0.254824
Lo Celso    0.000000
Messi       1.000000
Modric     -0.224502
Ramos      -0.695269
Suarez      0.700349
dtype: float64
```

En la salida por pantalla observamos la correlación entre el jugador Messi y el resto. Observamos que el jugador más similar a Messi es Suarez. ¿Tiene esto sentido? Pues bien, la similitud entre futbolistas depende de las valoraciones recibidas por los periodistas. Fijandonos en la matriz de puntuaciones inicial, el resultado obtenido es entendible pues por ejemplo, los periodistas Cristobal Soria y Lluis Flaquer, reconocidos forofos

del FCBarcelona, tienden a valor muy positivamente a estos jugadores frente a los jugadores del Real Madrid. Por contra, Tomas Roncero y Manolo Lama, fanáticos del Real Madrid tienden a valorar más positivamente a los jugadores blancos que a los culés. Por esta razón, observamos que la correlación existente entre Leo Messi y los jugadores de su club rival, Ramos y Modric, es inferior a 0 (la correlación es inversa), es decir, a valores altos de uno le suelen corresponder valor bajos del otro.

A continuación, mostramos en la siguiente tabla la similitud entre los diferentes futbolistas usando el coeficiente de Pearson como medida de similitud entre cada par:

	Banega	Lo Celso	Messi	Modric	Ramos	Suarez
Banega	1.000000	-0.866025	0.254824	-0.464286	0.000000	0.400066
Lo Celso	-0.866025	1.000000	0.000000	0.583212	0.654654	-0.707107
Messi	0.254824	0.000000	1.000000	-0.224502	-0.695269	0.700349
Modric	-0.464286	0.583212	-0.224502	1.000000	0.836538	-0.500000
Ramos	0.000000	0.654654	-0.695269	0.836538	1.000000	-0.575470
Suarez	0.400066	-0.707107	0.700349	-0.500000	-0.575470	1.000000

El coeficiente de Pearson se utiliza para examinar la fuerza y la dirección de la relación lineal entre dos variables continuas, de manera que mientras mayor sea el valor absoluto del mismo, más fuerte será la relación entre las variables. En nuestro caso, estas variables son cada uno de los futbolistas (items), de manera que una correlación cercana a 0 indica que no existe relación lineal entre los futbolistas; es el caso de los futbolistas Lo Celso y Messi. En relación a lo comentado tras los resultados obtenidos para el jugador Messi, observamos que el coeficiente de correlación entre Ramos y los jugadores del FCBarcelona, Messi y Suarez, es negativo indicando así que a aquellos periodistas que les gusten los jugadores del FCBarcelona probablemente no les guste el capitán del Real Madrid; por contra, posiblemente a aquellos periodistas que les guste Ramos también les gustará Modric ya que la correlación existente entre ellos es positiva.

4 Filtrado colaborativo

Dentro de los sistemas de recomendación basados en filtrado colaborativo, existen dos clasificaciones que son los **basados en memoria** y los **basados en modelos**. Nosotros nos centraremos en los métodos basados en memoria que emplean métricas de similaridad para determinar el parecido entre una pareja de usuarios. Para ello calcularemos los items que han sido votados por ambos usuarios y compararemos dichos votos para calcular la similaridad.

4.1 Similitud entre usuarios

Como comentamos en la identificación del problema, dos de las métricas más usuales para el cálculo de la similitud entre dos usuarios, es decir, la distancia entre usuarios son la distancia euclídea modificada y el coeficiente de Pearson. Como hemos visto en el análisis anterior, usaremos esta última para realizar nuestro cálculo de similaridad entre usuarios.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Mostramos en la siguiente tabla la similitud existente entre los usuarios (periodistas en nuestro caso) utilizando el coeficiente de Pearson:

	Cristobal Soria	Julio Maldonado	Lluis Flaquer	Manolo Lama	Manu Carreño	Tomas Roncero
Cristobal Soria	1.000000	0.188982	0.740779	-0.314270	0.374634	-0.114708
Julio Maldonado	0.188982	1.000000	0.578352	-0.705650	0.225668	-0.220755
Lluis Flaquer	0.740779	0.578352	1.000000	-0.839980	0.344381	-0.530105
Manolo Lama	-0.314270	-0.705650	-0.839980	1.000000	0.000000	0.800641
Manu Carreño	0.374634	0.225668	0.344381	0.000000	1.000000	-0.389249
Tomas Roncero	-0.114708	-0.220755	-0.530105	0.800641	-0.389249	1.000000

En la tabla adjunta observamos como los periodistas Lluis Flaquer y Cristobal Soria presentan una correlación positiva entre ellos y negativa con los periodistas Tomas Roncero y Manolo Lama. Dicha característica ya la habíamos comentado y es fruto de la adoración de los dos primeros por los jugadores culés y de los dos últimos por los merengues, así como el desprecio a los jugadores del máximo rival.

4.2 Cálculo de las valoraciones restantes

Una vez calculada la similitud entre cada par de usuarios (periodistas) nos disponemos a predecir las valoraciones que dichos usuarios no han realizado a determinados futbolistas. Hablaremos de dos formas de llevarlo a cabo:

1. **Aplicar la técnica de los *k*-Vecinos**: aplicando esta técnica ya conocida en el campo del Machine Learning, se obtienen los k periodistas más similares al periodista del cual se quiere conocer la valoración. Una vez calculadas todas las similaridades entre los periodistas, las ordenamos de mayor a menor para cada usuario y cogemos el número de vécinos que queramos (k) para posteriormente poder predecir la puntuación del usuario. Una vez que tenemos elegidos a los vecinos nos dispondremos a predecir las puntuaciones que un periodista realizaría sobre los jugadores que no ha valorado. Esta predicción se suele hacer con la media ponderada de los votos de los vecinos teniendo en cuenta la similaridad entre ellos.
2. Otra forma de llevar a cabo estas predicciones es tener en cuenta a todos los periodistas de nuestro set de datos, siempre y cuando hayan realizado la puntuación al jugador en cuestión, para realizar la predicción. De esta forma, similitudes más alta influirán más en la media ponderada.

Emplearemos la técnica de los k-Vecinos usando $k = 2$, es decir, usaremos los 2 periodistas más similares al que queremos conocer la valoración. Teniendo en cuenta la matriz de similitudes entre periodistas llegamos a la siguiente tabla donde observamos los 2 vecinos más cercanos a cada periodista, es decir, aquellos periodistas con mayor correlación al dado:

Periodista	Vecinos
Cistobal Soria	Lluis Flaquer, Manu Carreño
Julio Maldonado	Lluis Flaquer, Manu Carreño
Lluis Flaquer	Cristobal Soria, Julio Maldonado
Manolo Lama	Tomas Roncero, Manu Carreño
Manu Carreño	Cristobal Soria, Lluis Flaquer
Tomas Roncero	Manolo Lama, Cristobal Soria

Table 1: Técnica k-Vecinos

Nos encontramos ya en condiciones de empezar a realizar las predicciones. Mostraremos a continuación como se imputarían las valoraciones de Cristobal Soria a los jugadores Leo Messi y Lo Celso usando la técnica de

los k-Vecinos y utilizando todas las opiniones de los periodistas. Posteriormente, mostraremos en una tabla el resto de valoraciones predichas.

Analizamos en primer lugar la **predicción de la puntuación de Cristobal Soria a Leo Messi**:

- En la siguiente tabla se muestra la similitud de los distintos periodistas con Cristobal Soria y la puntuación que estos le han dado a Leo Messi:

```
corr_CS = pd.Series(corr_matrix_aux[0])
table_Messi = pd.concat([corr_CS, pd.Series(np.array(data['Messi']))], axis=1)
table_Messi = table_Messi.rename(columns = {0:'Pearson',1:'Messi'})
table_Messi.index = data.index
table_Messi
```

Periodistas	Pearson	Messi
Cristobal Soria	1.000000	NaN
Julio Maldonado	0.188982	5.0
Lluis Flaquer	0.740779	5.0
Manolo Lama	-0.314270	3.5
Manu Carreño	0.374634	3.0
Tomas Roncero	-0.114708	4.0

- **Utilizando la técnica de los k-Vecinos.** A continuación, eliminamos de la tabla anterior aquellas observaciones en las que Leo Messi no tiene puntuación (dichas observaciones no se utilizan en el cálculo de la media ponderada) e incorporamos una nueva columna que hace referencia a la puntuación ponderada. El próximo paso es seleccionar las dos observaciones (periodistas) con mayor similitud con nuestro periodista a predecir (2-Vecinos). Finalmente, imprimimos la valoración predicha en un comentario:

```
table_Messi_aux = table_Messi.dropna(how='any')
table_Messi_aux = table_Messi_aux.sort_values(by='Pearson', ascending=False)
table_Messi_aux = table_Messi_aux.iloc[:2,]
table_Messi_aux['Puntuacion ponderada'] = table_Messi_aux['Pearson'] * table_Messi_aux.Messi
table_Messi_aux
```

Periodistas	Pearson	Messi	Puntuacion ponderada
Lluis Flaquer	0.740779	5.0	3.703893
Manu Carreño	0.374634	3.0	1.123903

```
media_ponderada = sum(table_Messi_aux['Puntuacion ponderada'])/sum(table_Messi_aux.Pearson)
print("La valoración de Cristobal Soria sobre Leo Messi será de:", round(media_ponderada,1))
```

```
La valoración de Cristobal Soria sobre Leo Messi será de: 4.3
```

- **¿Qué puntuación habría recibido el item por el usuario si se hubieran considerado todas las opiniones de los periodistas?** Atendiendo al código no hay diferencias significativas pues se trata de considerar todas las opiniones de los periodistas en vez de quedarnos con los usuarios con mayor correlación al dado. En relación al resultado observamos que la valoración predicha es superior a la obtenida en la anterior técnica.

```
table_Messi_aux_complete = table_Messi.dropna(how='any')
table_Messi_aux_complete['Puntuacion ponderada'] = table_Messi_aux_complete['Pearson'] * table_Messi_aux_complete.Messi
media_ponderada_complete = sum(table_Messi_aux_complete['Puntuacion ponderada'])/sum(table_Messi_aux_complete.Pearson)
print("La valoración de Cristobal Soria sobre Leo Messi considerando las opiniones de todos los críticos es:",
      round(media_ponderada_complete, 1))
```

```
La valoración de Cristobal Soria sobre Leo Messi considerando las opiniones de todos los críticos es: 4.8
```

De manera análoga, realizamos la **predicción de la valoración de Cristobal Soria al futbolista Lo Celso**:

```
table_LoCelso = pd.concat([corr_CS, pd.Series(np.array(data['Lo Celso']))], axis=1)
table_LoCelso = table_LoCelso.rename(columns = {0:'Pearson',1:'Lo Celso'})
table_LoCelso.index = data.index
table_LoCelso
```

Periodistas	Pearson	Lo Celso
Cristobal Soria	1.000000	NaN
Julio Maldonado	0.188982	4.5
Lluis Flaquer	0.740779	3.5
Manolo Lama	-0.314270	NaN
Manu Carreño	0.374634	4.0
Tomas Roncero	-0.114708	4.0

- En este caso eliminaremos, además de la fila de índice Cristobal Soria (pues es el periodista del que queremos conocer la valoración hacia el futbolista en cuestión), la de Manolo Lama pues no tiene al periodista Lo Celso puntuado. Mostramos por pantalla la valoración estimada usando la **Técnica de los k-Vecinos** así como la alcanzada considerando las valoraciones dadas por todos los periodistas.

```
table_LC_aux = table_LoCelso.dropna(how='any')
table_LC_aux = table_LC_aux.sort_values(by='Pearson', ascending=False)
table_LC_aux = table_LC_aux.iloc[:2,]
table_LC_aux['Puntuacion ponderada'] = table_LC_aux['Pearson'] * table_LC_aux['Lo Celso']
table_LC_aux
```

Periodistas	Pearson	Lo Celso	Puntuacion ponderada
Lluis Flaquer	0.740779	3.5	2.592725
Manu Carreño	0.374634	4.0	1.498537

```
media_ponderada = sum(table_LC_aux['Puntuacion ponderada'])/sum(table_LC_aux.Pearson)
print("La valoración de Cristobal Soria sobre Lo Celso será de:", round(media_ponderada,1))
```

```
La valoración de Cristobal Soria sobre Lo Celso será de: 3.7
```

```
table_LC_aux_complete = table_LoCelso.dropna(how='any')
table_LC_aux_complete['Puntuacion ponderada'] = table_LC_aux_complete['Pearson'] * table_LC_aux_complete['Lo Celso']
media_ponderada_complete = sum(table_LC_aux_complete['Puntuacion ponderada'])/sum(table_LC_aux_complete.Pearson)
print("La valoración de Cristobal Soria sobre Lo Celso considerando las opiniones de todos los críticos es:",
      round(media_ponderada_complete,1))
```

```
La valoración de Cristobal Soria sobre Lo Celso considerando las opiniones de todos los críticos es: 3.8
```

El resto de puntuaciones predichas quedan reflejadas en la siguiente tabla, recordamos que había un total de 5 registros faltantes (faltan por imputar 3):

Periodista	Futbolista	Predicción k-Vecinos	Predicción considerando todos los periodistas
Julio Maldonado	Banega	3.4	3.7
Manolo Lama	Lo Celso	4	3.7
Tomas Roncero	Ramos	5	5

Table 2: Valoraciones predichas

Observamos que las valoraciones predichas son ligeramente distintas en función de la técnica que escojamos. A continuación mostraremos el set de datos donde los valores NaN han sido sustituidos por las predicciones alcanzadas por la ténica k-Vecinos.

Futbolistas / Periodistas	Banega	Lo Celso	Messi	Modric	Ramos	Suarez
Cristobal Soria	3.0	3.7	4.3	3.0	3.5	4.5
Julio Maldonado	3.4	4.5	5.0	3.5	3.0	3.5
Lluis Flaquer	3.5	3.5	5.0	1.0	2.0	5.0
Manolo Lama	3.5	4.0	3.5	4.5	5.0	3.5
Manu Carreño	3.0	4.0	3.0	2.0	3.5	3.0
Tomas Roncero	2.5	4.0	4.0	5.0	5.0	3.5

4.3 Realizar las recomendaciones

Recordamos que el fin de los sistemas de recomendación es proporcionar a los usuarios una serie de recomendaciones sobre un determinado tipo de items. En nuestro caso, se trata de asignar/recomendar a cada periodista una serie de jugadores. Así pues, una vez que se han obtenido todas las predicciones sobre la valoración de los jugadores que los periodistas no habían realizado, realizamos las recomendaciones; ordenamos las valoraciones de manera descendente y recomendamos aquellos jugadores con mayor valoración.

En la tabla adjunta aparecen los jugadores recomendados a los periodistas del filtrado colaborativo suponiendo que nuestro sistema de recomendación sólo recomienda 2 items.

Periodista	Futbolistas
Cristobal Soria	Suarez, Messi
Julio Maldonado	Messi, Lo Celso
Lluis Flaquer	Messi, Suarez
Manolo Lama	Ramos, Modric
Manu Carreño	Lo Celso, Ramos
Tomas Roncero	Modric, Ramos

Table 3: Recomendaciones