EXAMEN ACUE JUNIO 2017

Alumno: Moisés Frutos Plaza
Fecha: 14/06/2017

1. Recomendación de Tapas:

- Propuesta de dos Métodos de Predicción

El primer modelo que se me ha venido a la mente al ver el primer bloque de datos, donde sólo hay una tapa 't1', en un sólo restaurante 'r1', y en el que hay distintas valoraciones de distintos usuarios como carácter diferenciador, es un **modelo de regresión lineal** ya que estamos de hecho ante un problema de regresión que encaja con este tipo de modelo. Los únicos atributos con valores distintos son userID, globalRating, tapaRating; por lo que, siendo consecuentes con los datos que tenemos y el contexto del problema el único atributo realmente significativo y del que podemos hacer una predicción es sobre **globalRating**. De modo, que escojo este atributo como el más significativo.

En el segundo bloque de datos, observamos que todos los datos pertenecen a un mismo usuario, el usuario 'u31', apriroi lo más adecuado viendo los datos y el contexto sería hacer un **recomendador colaborativo basado en usuario**. Y éste es el que se propone como segundo modelo de predicción.

- Cálculos y comparativa da los dos modelos propuestos

Primer modelo: Modelo de Regresión Lineal

```
datos.examen <- read.csv("acue_examen_dataset.csv", header=TRUE)
str(datos.examen)

#Propuesta de Modelo de Regresión Lineal al haber una sóla variable
significativa
#siendo la tapa la misma en todo momento. Tomamos como variable principalmente
#significativa el rating global

#Hacemos la regresión lineal
regresion.lineal <- lm(tapaRating ~ globalRating, data = datos.examen)
summary(regresion.lineal)

Call:
lm(formula = tapaRating ~ globalRating, data = datos.examen)

Residuals:
     Min       1Q   Median       3Q      Max
-1.7063 -0.1987  0.2937  0.2937  1.3089

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1684     0.4481   4.839 1.57e-05 ***
globalRating   0.5076     0.0990   5.127 6.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.574 on 45 degrees of freedom
Multiple R-squared:  0.3687,    Adjusted R-squared:  0.3547
F-statistic: 26.29 on 1 and 45 DF,  p-value: 6.02e-06


#En los resultados del summary podemos ver que el Adjusted R-squared es de
0.3547
#lo cual no es que sea del todo bueno apriori, pero viendo el contexto del
problema
#y los datos que tenemos parece lo más adecuado. De modo que continuamos.

#Sacamos la media
media.global.rating <-
mean(datos.examen$globalRating[datos.examen$tapaID=="t1"])
media.global.rating

#Lo pasamos a dataframe
df = data.frame(globalRating = media.global.rating)
#Realizamos la predicción con el modelo de regresión lineal
prediccion.lineal <- predict(regresion.lineal, df)
prediccion.lineal
       1
4.390797
#La predicción es 4.390797
```

Segundo modelo: Modelo de Regresión Lineal

```
#La segunda propuesta es un modelo de recomendador colaborativo basado en
usuario

library(recommenderlab)
datos.examen.sin.u31 <- read.csv("acue_examen_sin_u31.csv", header=TRUE)
datos.examen.u31 <- read.csv("acue_examen_u31.csv", header=TRUE)

#Construimos la matriz poniendo como número de tapas el número de columnas y el
número
#de usuarios con el número de filas
matriz.datos <- matrix(NA, ncol=(nrow(datos.examen.u31)+1),
                       nrow=(nrow(datos.examen.sin.u31)+1))
matriz.datos[1,2:11] <- datos.examen.u31$tapaRating
matriz.datos[2:38,1] <- datos.examen.sin.u31$tapaRating

#Convertimos la matriz al tipo de datos legible por recommenderlab
md <- as(matriz.datos, "realRatingMatrix")

#Damos nombre tanto a las filas como a las columnas:
rownames(md)  <- c(unique(datos.examen.u31$userID),
unique(datos.examen.sin.u31$userID))
colnames(md) <- c(unique(datos.examen.u31$tapaID),
unique(datos.examen.sin.u31$tapaID))

#Ploteamos la imagen para ver que funciona
image(md)

#Creamos el recomendador basado en usuario que es el propuesto en un
#principio:
r1 <- Recommender(md, "UBCF") #Recomendador Colaborativo Basado en Usuario
p1 <- predict(r1, md[1], type="ratings")
as(p1,"list")
```

```
#No sé porqué pero no nos da una predicción tal y como esperaba, de hecho me
sale
#una salida extraña
$`1`
named numeric(0)


#Probamos con el recomendado basado en ítems
r2 <- Recommender(md, "IBCF") #Recomendador Colaborativo Basado en Ítem
p2 <- predict(r1, md[1], type="ratings")
as(p2,"list")
#Con este vuelve a pasar lo mismo
$`1`
named numeric(0)

#Probamos con el de Popularidad de los Ítems
r3 <-Recommender(md, method = "POPULAR") #Recomendador Basado en la popularidad
de los ítems
p3 <- predict(r3, md[1], n=5, type = "ratings")
prediccion.recomendador.popular <- as(p3, "list")
as.numeric(prediccion.recomendador.popular)

#Este ya sí, menos mal, nos da una predicción, la cual es de 4.4 que es muy
similar a la
#estimada por el modelo de regresión lineal 4.390797
```

Comparativa

```
diferencia <- abs(prediccion - as.numeric(prediccion.recomendador.popular))
diferencia
#La diferencia entre ambos modelos es de apenas 0.009202874. Es decir, muy
poca. Por lo
#tanto ambos modelos arrojan predicciones muy similares.
```

Como podemos apreciar en la última comparativa tanto el modelo de regresión lineal como el recomendador basado en la popularidad de los ítems tienen unas predicciones casi idénticas con una diferencia de apenas ***0.009202874*** lo cual es muy, muy poca la diferencia.

- Analizar método más apropiado

Yo personalmente viendo el conjunto de datos y el contexto del problema, en un principio pienso que un *recomendador basado en usuario* sería lo mejor para dar una recomendación al usuario 31, pero después de realizar los cálculos y las evaluaciones está claro que me quedaría con el **modelo de regresión lineal** ya que creo que es el que más se ajusta a la realidad del problema. Aunque la predicción del modelo basado en la popularidad de los ítems también da un resultado muy similar, yo creo que el modelo de regresión lineal es más preciso en este aspecto, pese a su bajo Adjusted R-squared.

- Estrategia o Modelo de Recomendación más adecuado para resolver el problema

Los datos que tenemos son escasos una buena estrategia sería recopilar más datos para poder realizar un mejor entrenamiento de los modelos, porque por ejemplo en el caso del modelo de regresión lineal el coeficiente de regresión es bastante bajo, se podría tratar de

aumentar ese coeficiente aportando más datos al conjunto. Si obtuviésemos un coeficiente de regresión lo suficientemente alto 0.8 o más, yo creo que sería el modelo idóneo para este problema, pero lo dicho, necesitaríamos recopilar más datos.

2. Empresas que han pasado por el ciclo de conferencias

- Conferencia HP

- **Big Data:** Recuerdo algo de que nos comentaron que trabajaban con Spark, para trabajar con Spark necesitas trabajar internamente sobre HDFS de Hadoop, y esta herramienta de Apache está totalmente orientada al Big Data.

- **Data Science:** Intentaban extraer conocimiento a partir de los datos para intentar predecir cuando iba a fallar una máquina.

- **Data Mining:** Exploraban y analizaban los datos con R y con Python. Tuvimos una pequeña charla sobre cual de los dos lenguajes era mejor. A mí personalmente me gusta más R, porque me siento más cómodo a la hora de realizar los cálculos, pero reconozco que visualmente Python junto con Scikit-Learn y Jupyter-Notebook pueden ofrecer una herramienta extra de presentación que se podría utilizar por ejemplo precisamente para una conferencia o explicar algo a la dirección de la organización.

- Conferencia Bahía

- **Big Data:** Recuerdo que utilizaban Hadoop y como base de datos ArangoDB, que es una base de datos NoSQL que combina documentos y grafos. El doble uso es muy bueno debido a que por un lado coleccionas los documentos, y por el otro puedes formar grafos para mejorar el tiempo de respuesta en las consultas. Esto sobre todo es muy útil cuando tienes muchísimos datos.

- **Data Science:** Se servían de unos principios fundamentales para intentar extraer información a partir de los datos pero no recuerdo cuales eran.

- **Data Mining:** Recuerdo que partían de informes médicos para hacer minería de datos, pero no recuerdo que tecnologías utilizaban para .

3. Respuestas breves

- **Big Data** es un concepto que engloba la necesidad de almacenar un gran volumen datos, la capacidad de ser capaces de procesarlos al tiempo que se producen, y al mismo tiempo intentar extraer conocimiento a partir de

ellos de modo que nos puedan dar una ventaja competitiva frente a alguien que no lo tenga. el volumen era su gran reto. **Data Science** es una rama dentro del Big Data que precisamente se encarga de ésto último, de buscar modelos que puedan predecir o extraer información a partir de un conjunto de datos dado, en el que por sí mismos no nos revelan apenas nada, o nada directamente.

- Si la NASA quiere enviar una transbordador a la Luna, primero tendrá que calcular la trayectoria de dicho transbordador y aparte tiene que predecir que en su trayecto no vaya a colisionar ningún objeto de los denominados basura espacial que están en la órbita terrestre. La NASA es más que probable que cuente con los datos de las trayectorias de todos estos objetos y pueda calcular la trayectoria del transbordador de modo que así evite una colisión que en el vacío del espacio puede resultar fatal para la nave y su tripulación.

- La utilidad es una valoración numérica de las preferencias de un usuario respecto a una alternativa dada. O lo que es lo mismo, la cualidad que tiene para alguien el que le sea útil o no una alternativa. Para la toma de decisiones se puede cuantificar la utilidad esperada con el resultado de cada decisión. El Principio MEU (Maximum Expected Utility) nos indica que debemos quedarnos con el máximo de utilidad esperada para las alternativas calculadas û(c,a), donde 'c' representa al usuario y 'a' a la aternativa, y 'û' su utilidad esperada (Acompaño en el script FuncionesACUE.R el ejercicio donde pudimos ver esto en clase).

- El resultado a bote pronto es del 41.37931%. Pero lo bueno del ejercicio es cómo hemos llegado hasta él. Y hemos llegado utilizando el Teorema de Bayes (dejo los cálculos hechos con más rigor en el script FuncionesACUE.R). Por un lado tenemos la hipótesis de que el coche sea azúl con 80% de probabilidad dicho por el testigo, y por otro tenemos la evidencia de que el 15% de los taxis de la ciudad son azules. Por el contrario, tenemos la hipótesis de que el coche sea verde con 20%, y la evidencia de que los taxis verdes de la ciudad es del 85%. Si usamos el Teorema de Bayes:

$$P(E|H)=\frac{P(H|E)\cdot P(H)}{P(E)}=\frac{0.80\cdot 0.15}{0.29}=0.41$$