

Proyecto TID

Alejandro Campoy Nieves

Gema Correa Fernández

Luis Gallego Quero

Jonathan Martín Valera

Andrea Morales Garzón

14 de noviembre de 2018

Contents

1. Comprender el problema a resolver	3
2. Preprocesamiento de datos	3
2.1. Lectura de datos	3
2.2. Falta de datos, categorización, normalización, reducción de dimensionalidad.	4

1. Comprender el problema a resolver

Para la realización y aplicación de las técnicas explicadas a lo largo del curso, hemos seleccionado un *dataset* proporcionado por *UCI Machine Learning Repository*. En concreto, hemos escogido **Drug Review Dataset**, una exhaustiva base de datos de medicamentos organizada por relevancia para medicamentos específicos. El conjunto de datos proporciona revisiones de pacientes sobre medicamentos específicos junto con las condiciones relacionadas. Además, las revisiones se agrupan en informes sobre tres aspectos: beneficios, efectos secundarios y comentarios generales. De igual modo, las calificaciones están disponibles con respecto a la satisfacción general, así como una calificación de efectos secundarios de 5 pasos y una calificación de eficacia de 5 pasos. Los datos se obtuvieron rastreando los sitios de revisión farmacéutica en línea.

DataSet		Number of			
Characteristics:	Multivariate, Text	Instances:	4143	Area:	N/A
Attribute	Integer	Number of	8	Date Donated	2018-10-
Characteristics:		Attributes:			02
Associated Tasks:	Classification, Regression, Clustering	Missing Values?	N/A	Number of Web Hits:	7001

Los datos se dividen en un conjunto train (75%) y otro conjunto test (25%) y se almacenan en dos archivos.tsv (tab-separated-values), respectivamente. Los atributos que tenemos en este dataset son:

1. **urlDrugName** (categorical): nombre de la droga
2. **condition** (categorical): nombre de la condición
3. **benefitsReview** (text): paciente sobre beneficios
4. **sideEffectsReview** (text): paciente sobre los efectos secundarios
5. **commentsReview** (text): comentario general del paciente
6. **rating** (numerical): clasificación de paciente de 10 estrellas
7. **sideEffects** (categorical): clasificación de 5 pasos de efectos secundarios
8. **effectiveness** (categorical): clasificación de efectividad de 5 pasos

2. Preprocesamiento de datos

Para poder analizar el dataset y realizar el preprocesamiento al mismo, lo primero que se va a hacer es leer tanto el conjunto de datos train como de test. Primero, leeremos los datos con los que se va a entrenar y luego los datos test.

2.1. Lectura de datos

A continuación, leemos nuestro dataset train y test:

```
# Lectura de datos train
datos_train <- read.table("datos/drugLibTrain_raw.tsv", sep="\t", comment.char="",
                          quote = "\"", header=TRUE)
head(datos_train, 5) # visualizar las 5 primeras filas
summary(datos_train) # información sobre los datos
View(datos_train)    # vista de la tabla

# Lectura de datos test
datos_test <- read.table("./datos/drugLibTest_raw.tsv", sep="\t", comment.char="",
```

```
                                quote = "\"", header=TRUE)
head(datos_test, 5) # visualizar las 5 primeras filas
summary(datos_test) # información sobre los datos
View(datos_test)    # vista de la tabla
```

2.2. Falta de datos, categorización, normalización, reducción de dimensionalidad.