Universidad del Rosario, Facultad de Economía
Big Data: Machine Learning
Taller 3

March 11, 2017

1. **Árboles de Decisión**

   El propósito de este ejercicio es predecir presencia de células cancerosas usando árboles de decisión, y en específico, el algoritmo C5.0. Para ello, usaremos la base `wisc_bc_data.csv` que utilizamos en el taller anterior.

   (a) Cargue la base en un data frame llamado `cancer`. Inspeccione la base con la función `str()`. ¿Cuántas observaciones tiene la base? ¿Qué nombre tiene el outcome que nos interesa predecir?

   (b) Limpie los datos para poder implementar el algoritmo. Asegúrese de eliminar la variable de identificación. Además, convierta el outcome de interés en un factor. Por medio de una tabla de proporciones, determine qué porporción de los casos son benignos y malignos.

   (c) Divida la base de datos en dos: entrenamiento y prueba. Para esto, aleatorice las filas que formarán parte de la base de entrenamiento y las de prueba. Fije la semilla en 123. Llame a sus bases de entrenamiento y prueba `cancer_train` y `cancer_test`. Inspeccione estas dos bases con la función `str()` y construya tablas de proporción para los outcomes en cada base. ¿Encuentra coherencia con las proporciones de la base original?

   (d) Utilizando el paquete `C50` construya un árbol de decisión. ¿Cómo es este árbol? Represéntelo gráficamente (es pequeño). De acuerdo con los resultados del árbol, ¿cuáles son las características de las biopsias más importantes al momento construir los subgrupos? ¿Qué porcentaje de precisión tiene el algoritmo sobre la base de entrenamiento?

   (e) Nuevamente, usando el paquete `C50` realice la predicción sobre la base de prueba. Construya una tabla en la que represente en el eje X el verdadero diagnóstico de cada biopsia y en el eje Y las predicciones del modelo. ¿Qué tan preciso es el modelo? ¿Cuál es la proporción de falsos positivos? ¿De falsos negativos?

2. **Naive Bayes**

   En este ejercicio usaremos el algoritmo Naive Bayes para predecir el partido político (Demócrata o Repúblicano) al que pertenecen representantes a la Cámara en Estados Unidos. Para ello, usaremos la base `votos84.csv` que contiene información sobre un conjunto de representantes a esta corporación en 1984. También se cuenta con información sobre cómo votaron (Sí o No) estos representantes frente a 16 proposiciones, durante su periodo legislativo.

   (a) Cargue la base en un data frame llamado `votos84`. Inspeccione la base con la función `str()`. ¿Cuántos representantes a la Cámara componen la base? ¿Cuántas variable tiene la base? ¿Cuál de ellas es el outcome de interés?

(b) Haga una gráfica de barras que indique el número de representantes que votan a favor y en contra de la proposición 1. Realice la misma gráfica, pero solo para representantes demócratas. Haga lo mismo solo para los republicanos. ¿Estos patrones de votación le dicen algo sobre la probabilidad de ser demócrata o republicano en función de cómo se vota por esta proposición?

(c) Divida la base de datos en dos: entrenamiento y prueba. Para esto, aleatorice las filas que formarán parte de la base de entrenamiento y las de prueba. Fije la semilla en 123. Llame a sus bases de entrenamiento y prueba `votos84_train` y `votos84_test`, respectivamente. Inspeccione estas dos bases con la función `str()`.

(d) Construya el par de vectores de outcomes y llámelos `votos84_train_labels` y `votos84_test_labels`. Utilice las bases que creó en el ítem anterior para ello.

(e) Limpie las bases de entrenamiento y prueba `votos84_train` y `votos84_test`, eliminando el outcome de interés (primera columna). Revise con la función `str()` que dicha eliminación sea acertada.

(f) Usando el paquete `e1071`, entrene un modelo por medio del algoritmo Naive Bayes. Utilice este mismo paquete para hacer las predicciones en la base de prueba. Construya una tabla en la que represente en el eje X el verdadero partido de cada representante y en el eje Y las predicciones del modelo. ¿Qué tan preciso es el modelo? ¿Cuál es la proporción de demócratas incorrectamente clasificados? ¿De republicanos?

(g) Estime un modelo alternativo, esta vez utilizando el estimador de Laplace, haciendo dicho parámetro igual a 1. ¿Mejoran las predicciones con este nuevo modelo? ¿Por qué sí o por qué no?

3. **Análisis de Texto**

La base de datos `tweets.csv` (https://goo.gl/CVpOeR) contiene 100 tweets recolectados durante los meses previos a octubre de 2016. Cada una de las observaciones corresponde a un tweet que habla sobre temas concernientes con el proceso de paz colombiano. La variable `sentimiento` está codificada de la siguiente manera: 1 cuando el sentimiento del tweet es positivo y 0 cuando el sentimiento del tweet es negativo. Utilizando sus conocimientos de R:

(a) Genere un *corpus* que contenga todos los tweets.

(b) Limpie cada tweet quitando: números, signos de puntuación, espacios y símbolos.

(c) Haga el top 10 de las palabras más frecuentes utilizadas en los tweets.

(d) Limpie cada tweet convirtiendo todo el texto a minúscula, quitando *stopwords* y realizando *stemming*.

(e) Haga el top 10 de las palabras más frecuentes utilizadas en los tweets.

(f) Explique por qué existen diferencias entre ambos rankings de frecuencias. ¿Cuál es más confiable? Explique su respuesta.

(g) Genere una nube de palabras para los tweets con sentimiento = 1 y otra para los tweets con sentimiento = -1.

(h) ¿Hay diferencia en el lenguaje según el sentimiento? Explique su respuesta.

(i) Haciendo la limpieza respectiva, encuentre cuáles son las 5 palabras que se repiten más veces en las biografías de los usuarios que hacen tweets negativos sobre el proceso de paz.

Fecha de entrega: sábado 18 de marzo. Enviar a jdavidmartinezg@gmail.com