Caso de estudio: “Titanic”

Probabilidad condicional, Independencia y Teorema de Bayes

Introducción [5 min]

El Royal Mail Ship Titanic fue un transatlántico británico, el mayor barco de pasajeros del mundo al finalizar su construcción, que se hundió en la noche del 14 a la madrugada del 15 de abril de 1912 durante su viaje inaugural desde Southampton a Nueva York. En el hundimiento del Titanic murieron 1496 personas de las 2208 que iban a bordo, lo que convierte a esta tragedia en uno de los mayores naufragios de la historia ocurridos en tiempo de paz. Construido entre 1909 y 1912 en los astilleros de Harland & Wolff en Belfast, el Titanic era el segundo de los tres buques que formaban la clase Olympic, propiedad de la naviera White Star Line, junto al RMS Olympic y, posteriormente, el HMHS Britannic. [Wikipedia]

Por otro lado, existe una plataforma llamada Kaggle que sirve para realizar competencias de aprendizaje automático (Machine Learning) en la cual se dispone de datasets (tablas de datos) para generar modelos que puedan predecir alguna cosa. En este caso, estudiaremos los datos de la competencia del Titanic. Estas competencias suelen tener dos tablas, una es utilizada para entrenamiento el modelo y otra para validarlo. La segunda tabla presenta datos nuevos respecto a los de entrenamiento con el fin de evaluar el modelo en condiciones nuevas y analizar cómo performa. En este caso, no nos interesa generar ningún modelo sino que nos interesa analizar la información y utilizar con datos reales los conceptos estudiados en clase.

Conceptos previos

- Probabilidad condicional
- Independencia de eventos
- Teorema de Bayes

R y RStudio [25min]

¿Qué es R?

Pueden ver una descripción más extensa aquí. R es un lenguaje de programación que provee muchísimas facilidades para la manipulación y trabajo con datos de diversos tipos para realizar trabajos estadísticos. Nace como sucesor de S, otro lenguaje de programación focalizado en aplicaciones estadísticas.
Existe un ecosistema de librerías, un montón, para muchísimas aplicaciones lo cual permite que la reutilización y aprovechamiento de herramientas sea muy propicio para el desarrollo de estudios e investigaciones además de aplicaciones industriales (profesionales).

¿Qué es R Studio?

Es un programa que permite facilitar

¿Cómo lo instalo?

Pueden visitar esta página que tiene servidores de descarga en múltiples países (en Argentina reside en la Universidad de La Plata)

¿Qué necesito saber?

Nada a priori. Vamos a mostrar ejemplos de código que les van a servir como herramienta de verificación. Es totalmente optativo y pueden obviarlo por completo. Sin embargo, si les interesa, pueden revisar la página de manuales para tener guías formales de referencia y de libros.

Código en R [15min]

```
rm(list = ls())

install.packages("titanic")

library(titanic)

datos <- titanic_train

# Imprimo las primeras 7 filas de la tabla
head(datos, n = 7)

# Imprimo las ultimas 8 filas de la tabla
tail(datos, n = 8)

# Evaluo las columnas disponibles en la tabla
str(datos)

# Cuantas filas tengo?
nrow(datos)

# Cuantas columnas tengo?
ncol(datos)
length(datos)

# Cuanta gente se salvo en el data set de train?
# Nota: 0 significa que no y 1 significa que si.
table(datos$Survived)
# Cual es la proporcion de gente que se salvo?
```

```
prop.table(table(datos$Survived))

# Como es la composicion de la gente que se subio?
table(datos$Sex)
# Existen datos faltantes?
sum(is.na(datos$Sex))
# Y como es la proporcion de gente que se salvo vs. el sexo?
table(datos$Sex, datos$Survived)
prop.table(table(datos$Sex, datos$Survived))

# Como es la composicion de clases de tickets?
table(datos$Pclass)
prop.table(table(datos$Pclass))
# Y versus si sobreviven?
table(datos$Pclass, datos$Survived)
prop.table(table(datos$Pclass, datos$Survived))
# Existen datos faltantes?
sum(is.na(datos$Pclass))

# Edad. Miremos un poco...
table(datos$Age)
# No se ve muy bien verdad? Ajustemos de a bloques de a 10 años
datos$AgeDec <- as.integer(datos$Age / 10)
table(datos$AgeDec)
# Noten que no tenemos datos en varios
sum(is.na(datos$AgeDec))
# Y versus si sobreviven?
table(datos$AgeDec, datos$Survived)
prop.table(table(datos$AgeDec, datos$Survived))

# Donde embarcaron?
unique(datos$Embarked)
table(datos$Embarked)
prop.table(table(datos$Embarked))
# Hay relacion entre si embarcaron y si sobrevivieron?
table(datos$Embarked, datos$Survived)
prop.table(table(datos$Embarked, datos$Survived))
# Tenemos datos faltantes?
sum(is.na(datos$Embarked))

# Tablas de mas de dos propiedades
table(datos$AgeDec, datos$Sex, datos$Survived)
```

Analizando el dataset

¿Qué información dispongo? [5min]

Descripción de (algunas) las columnas:

Columna	Tipo de dato y referencias
PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...	Entero.
$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...	Lógico. 0 → No sobrevivió, 1 → sobrevivió
$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...	Entero. 1,2,3 clase.
$ Sex : chr "male" "female" "female" "female" ...	Categórico. female → mujer, male → hombre
$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...	Entero: edad.
Embarked : chr "S" "C" "S" "S" ...	Categórico. C = Cherbourg, Q = Queenstown, S = Southampton

Tablas de contención [2min]

Las tablas de contención son tablas en donde podemos agregar dos o más categorías de forma sintética y que expresan el resultado de una sucesión de observaciones. En este caso, disponemos de las siguientes tablas que van a permitir resolver preguntas alrededor de la parte de la tripulación que no disponemos.

Total de pasajeros en el dataset
891

Puerto	Sobrevivientes
NS/NC	2
Cherbourg	93
Queenstown	30
Southampton	217

Hombres que sobrevivieron	109
Hombres que no sobrevivieron	468

Primera clase	**Segunda clase**	**Tercera clase**
216	184	491

Primera clase que sobrevivió	**Segunda clase que sobrevivió**	**Tercera clase que sobrevivió**
136	87	119

Rango etario	**No sobrevivió**		**Sobrevivió**	
	Mujer	**Hombre**	**Mujer**	**Hombre**
0-9 años	11	13	19	9
10-19 años	11	50	34	7
20-29 años	20	123	52	25
30-39 años	10	84	50	23
40-49 años	10	45	22	12
50-59 años	2	26	16	4
60-69 años	0	13	4	2
70-79 años	0	6	0	0
80-89 años	0	0	0	1

Ejercicios [40min]

1. De la gente que sobrevivió, se dispone la información de donde embarcó. Sabiendo que el total de la muestra son 891 pasajeros, ¿cuál es la proporción de sobrevivientes?
2. ¿Y cuántos no sobrevivieron?
3. Dados los sucesos *A*: *"sobrevivió al accidente del Titanic"* y *B: "es mujer"* determinar si los eventos A y B son independientes sabiendo que la cantidad de personas que sobrevivió o que son mujeres son 423.

4. Dado una nueva forma de repartir el espacio muestral a partir los eventos {“tripulante de 1ra clase”, “tripulante de 2da clase”, “tripulante de 3ra clase”}, ¿cuál es la probabilidad que un tripulante que no sobrevivió sea de primera, segunda o tercera clase?
5. A partir de la información de de rangos etarios, computar la probabilidad que:
    a. un hombre mayor de 30 años haya sobrevivido,
    b. dado que una mujer sobrevivió tenga entre 10 y 50 años,
    c. dado que no sobrevivió una persona, tenga entre 0 y 20 años.
    d. dado que no sobrevivió una persona, sea mujer y tenga más de 50 años.