Machine Learning para el análisis de texto

Daniel Jiménez M.

Universidad Nacional de Colombia

16 -11 -2020

Bibliografía

- Brett Lantz, Machine Learning with R, 2016, PACKT
- Bradley B & Brandon G, Hands-On Machine Learning with R,2020, The R Series.
- Silge G, Kuhn M,Tidy Modeling with R,2020,The R Series.
- Silge G,Hvitfeldt E, Supervised Machine Learning for Text Analysis in R, 2020, The R Series.

Librerías

Las librerías para esta clase son :

```
library(tidyverse)
library(tidymodels)
library(skimr)
```

Algunas definiciones importantes

La ciencia de datos, el Machine Learning y la AI son areas diferentes de estudio y trabajo, a pesar que se ha popularizado el termino **Data Scientist** como una sumatoria de los anteriores y por fines técnicos esto se mantendrá así. Más a continuación se presentarán algunas definiciones importantes y el contexto en el cual se desenvuelven.

Algunas definiciones importantes

La siguiente imagen es una tesis de Robinson (2017) sobre las diferencias en acciones de los data scientist:

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

Data Science

El foco principal de la ciencia de datos es la producción de insigths, esto se logra de la siguiente manera :

- Los insigths son entendibles ;
- Se basan en el entendimiento descriptivo y prescriptivo de un problema;
- Se intenta entender las causalidades entre las variables exploratorias;

Una definición clásica dice que el data science combina : Estadística, ingeniería de sistemas y conocimiento de negocios.

Data Science

Pero a diferencia de la definición clásica que es la que proveé al nombre de Data Scientist a la persona que hace un poco de todo, es la siguiente:

Data Science

El data science es la disciplina que envuelve las siguientes corrientes:

- Inferencia estadística;
- Visualización de datos;
- Diseño de experimentos;
- Dominio del negocio y lo más importante;
- Comunicación.

Machine Learning

El machine learning produce predicciones (llamese forecast o clasificación, detección de anomalías entre otros).

Se divide en :

- Análisis supervisado: los datos tienen la forma de $\{(x_1, y_1), ..., (x_n, y_n)\}$
  - El objetivo es estudiar el comportamiento de la variable **y**, condicional a las variables respuestas.
  - Puntualmente se debe entender el comportamiento (distribución) de la variable **y** dada las explicativas.

Machine Learning

- Análisis no supervisado: Los datos tiene la siguiente forma $\{x_1, ...x_n\}$
  - El objetivo es estudiar la variable **x** y los posibles conglomerados que se encuentren en ello.
  - Matemáticamente hay que estudiar la variable **x**

Inteligencia Artificial

Son acciones, desarrolladas por algoritmos que entienden la lógica del funcionamiento de un entorno.

Una inteligencia artificial que todos usamos (o eso espero) es google maps, cuando selecciona la mejor ruta!

Machine Learning para la clasificación de textos