

EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2023 — Introducción.

Objetivo. Dar un panorama de lo que rige los principios del curso de aprendizaje estadístico y diferenciarlo de otras estrategias de modelado predictivo. Sentar la notación base.

1. INTRODUCCIÓN

Herramientas para entender reglas de asociación. Con el objetivo de generar predicciones acertadas. No es estadística. No es inferencia causal.

1.1. Características del Aprendizaje Estadístico

- Flexibilidad.
- Procesamiento automático.
- Complejidad (en datos).
- Predicción.

En aprendizaje estadístico usualmente consideramos menos suposiciones de los datos en contraste con estadística; pensamos que el procesamiento es automático; que los datos son complejos; y que el interés primordial es la **predicción**.

1.2. Filosofía del curso

Es importante entender los modelos, la intuición y fortalezas y debilidades de los métodos que se utilizan en tareas de aprendizaje **supervisadas** y no **supervisadas**.

1.3. Principios

Consideraremos los siguientes, tomados de [1].

- Muchos métodos de aprendizaje son relevantes para una gran variedad de aplicaciones.

No encasillar en el ámbito académico o estadístico. Por supuesto hay muchos mas modelos de los que veremos pero nos concentraremos en los que no son de nicho.

- Aprendizaje estadístico **no** es una colección de cajas negras.

Lamentablemente no hay un método que sea exitoso para cualquier tipo de aplicación. Tenemos que conocer bien nuestras herramientas para saber cuál usar en qué situación.

- Una cosa es entender cómo funciona; otra, implementarlo desde cero.

No reinventaremos la rueda. En el curso nos concentraremos en las ideas, no en la implementación.

- Conocimiento de dominio. Obtención de información relevante.

Espero poder transmitir la necesidad de pensar en la aplicación del modelo. Algo que no se ve dentro de una formulación matemática. Pero la cual, si no se es cuidadoso podría tener en consecuencia una herramienta inservible al mediano/largo plazo.

2. DISTINCIOS CON RESPECTO A APRENDIZAJE DE MÁQUINA (ML)

En aprendizaje estadístico nos interesa comprender el proceso que genera los datos y su representatividad estadística.

3. MÉTODOS

Aprendizaje supervisado y aprendizaje no supervisado.

3.1. Tareas de predicción

- Clasificación
- Regresión

4. NOTACIÓN

Denotamos por x una **variable aleatoria** y por $\mathbb{P}(\cdot)$ una **función de distribución**. Escribimos $x \sim \mathbb{P}$ para denotar que la variable aleatoria x tiene distribución $\mathbb{P}(\cdot)$. Denotamos por $\mathbb{E}[\cdot]$ el **valor esperado** del argumento con respecto a la distribución que estamos considerando. Durante el curso seremos explícitos en la variable aleatoria y usaremos

$$\mathbb{E}_x[\cdot] = \int_{\mathcal{X}} \cdot \pi(x) dx, \quad (1)$$

o bien, haremos énfasis en la distribución por medio de lo siguiente

$$\mathbb{E}_{\pi}[\cdot] = \int_{\mathcal{X}} \cdot \pi(x) dx, \quad (2)$$

de acuerdo al contexto.

Denotamos por n el **número de observaciones**; p para el **número de características** de dichas observaciones. Así que, x_{ij} con $i = 1, \dots, n$ y $j = 1, \dots, p$ será un elemento de nuestras observaciones.

4.1. Tipos de características

4.2. Notación: objetivos y muestras

Usualmente tendremos una característica que queremos predecir y la denotamos por y . Consideraremos $y \in \mathbb{R}$ ó $y \in \{0, 1\}$ ó $y \in \{1, 2, \dots, K\}$.

El conjunto de datos que tenemos **disponible para entrenar** modelos lo denotamos por

$$\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}. \quad (3)$$

Es ideal considerar que además tenemos datos adicionales para **hacer pruebas**. A este conjunto lo denotaremos por

$$\mathcal{T}_m = \{(x_1, y_1), \dots, (x_m, y_m)\}. \quad (4)$$

Possiblemente necesitemos notación mas especializada para hacer distinciones adicionales o el contexto nos ayude a requerir una notación mas laxa. Esto lo definiremos sobre la marcha.

5. REPASO DE PROBABILIDAD

Consideraremos como requisitos el contenido de Cálculo de Probabilidades II y Álgebra Lineal (o equivalentes). En particular lo que requerimos como base es lo siguiente.

5.0.1. Definición [Espacio de Probabilidad]: Un espacio de probabilidad está definido por la terna $(\Omega, \mathcal{X}, \mathbb{P})$:

1. El espacio muestral, Ω (elementos).
2. El espacio de eventos medibles, \mathcal{X} (subconjuntos).
3. La medida de probabilidad, $\mathbb{P} : \mathcal{X} \rightarrow [0, 1]$.

5.0.2. Definición [Variable aleatoria]: Una variable aleatoria es una función $X : \mathcal{X} \rightarrow \mathbb{R}$ con la propiedad de que las pre-imágenes bajo X son eventos medibles. Es decir,

$$\{w \in \mathcal{X} : X(w) \leq x\} \in \mathcal{X} \quad \forall x \in \mathbb{R}. \quad (5)$$

5.0.3. Definición [Función de acumulación]: Para toda variable aleatoria X tenemos una función de acumulación $\mathbb{P}_X : \mathbb{R} \rightarrow [0, 1]$ dada por

$$\mathbb{P}_X(x) = \mathbb{P}(\{w \in \mathcal{X} : X(w) \leq x\}). \quad (6)$$

Esto usualmente lo escribimos como $\mathbb{P}_X(x) = \mathbb{P}\{X \leq x\}$.

5.0.4. Definición [Función de densidad]: Una variable aleatoria es continua si su función de acumulación es absolutamente continua y puede ser expresada por medio de

$$\mathbb{P}_X(x) = \int_{-\infty}^x \pi(s) ds, \quad (7)$$

donde la anti-derivada $\pi : \mathbb{R} \rightarrow [0, \infty)$ se llama la **función de densidad** de la variable aleatoria X .

Las propiedades generales de las distribuciones de probabilidad se pueden especificar por medio de su centralidad (localización), su dispersión, su rango de valores, su simetría y el comportamiento de valores extremos.

En general esto lo podemos extraer de los momentos

$$\mathbb{E}(X^p) = \int_{\mathbb{R}} x^p \pi(x) dx, \quad (8)$$

o los momentos centrales. Por ejemplo: media y varianza.

Uno de los resultados que espero recuerden bien de sus cursos anteriores es el de la **Ley de los Grandes Números**. La cual podemos enunciar como:

5.0.5. Teorema [Ley de los Grandes Números]: Sea X_1, X_2, \dots una colección de variables aleatorias independientes e idénticamente distribuidas (iid) y sea \bar{X}_n el promedio de un subconjunto de n . Si denotamos por μ el valor promedio de X_i dentro de esa colección, entonces tenemos que

$$\bar{X}_n \rightarrow \mu \quad (\text{casi seguramente}). \quad (9)$$

5.0.6. Teorema [Límite Central]: Sea X_1, \dots, X_n una colección de n variables aleatorias iid con $\mathbb{E}[X_i] = \mu$ y $\mathbb{V}[X_i] = \sigma^2 < \infty$. Entonces

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad (10)$$

para n suficientemente grande.

6. CONTROL DE VERSIONES

Los *softwares de control de versiones* nos permiten llevar un registro y administración de cambios en archivos. Usualmente para proyectos de programación.

Ayudan a trabajar colaborativamente en ambientes de equipos de trabajo.

Aunque no exploraremos *todo* lo que se puede hacer con **Git** y **GitHub** lo usaremos para llevar un control del desarrollo y de entrega de tareas. Usaremos los principios mas básicos.

7. R STATISTICAL PROGRAMMING LANGUAGE

R es un lenguaje de programación orientado a cómputo estadístico y generación de gráficos estadísticos. Está escrito para interactuar por medio de ejecución de *scripts* (archivos de texto con instrucciones) o la consola interactiva. Ver Fig. 1.

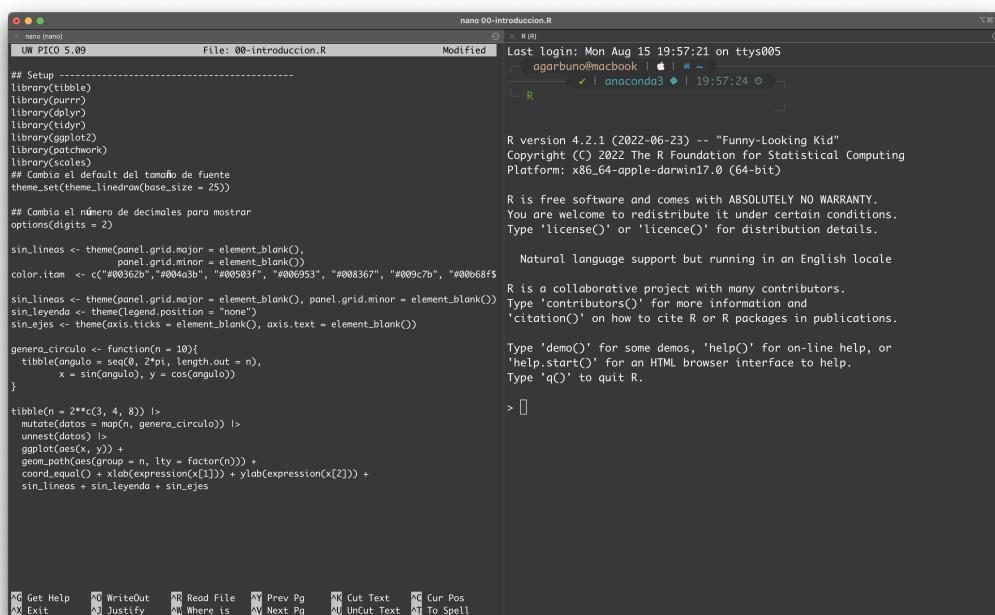


FIGURA 1. Dos ventanas, un editor de texto y una consola de R.

Es usual utilizar un ambiente de desarrollo para programar e interactuar con el lenguaje. Para R el mas común es **Rstudio** el cual tiene además algunas extensiones útiles para el desarrollo de análisis estadístico. Ver Fig. 2.

Visual Studio Code es una alternativa multi-lenguaje para desarrollar proyectos de análisis estadístico en R. Ver Fig. 3.

Y habemos los que nos *conformamos* con un buen editor de texto como ambiente de desarrollo. Ver Fig. 4.

8. AMBIENTE DE R

```

1 library(tidyverse)    # Herramientas de procesamiento
2 library(tidymodels)   # Herramientas de modelado
3 library(ISLR)         # Datos del libro de texto
4 library(MASS)         # Datos de Boston

```

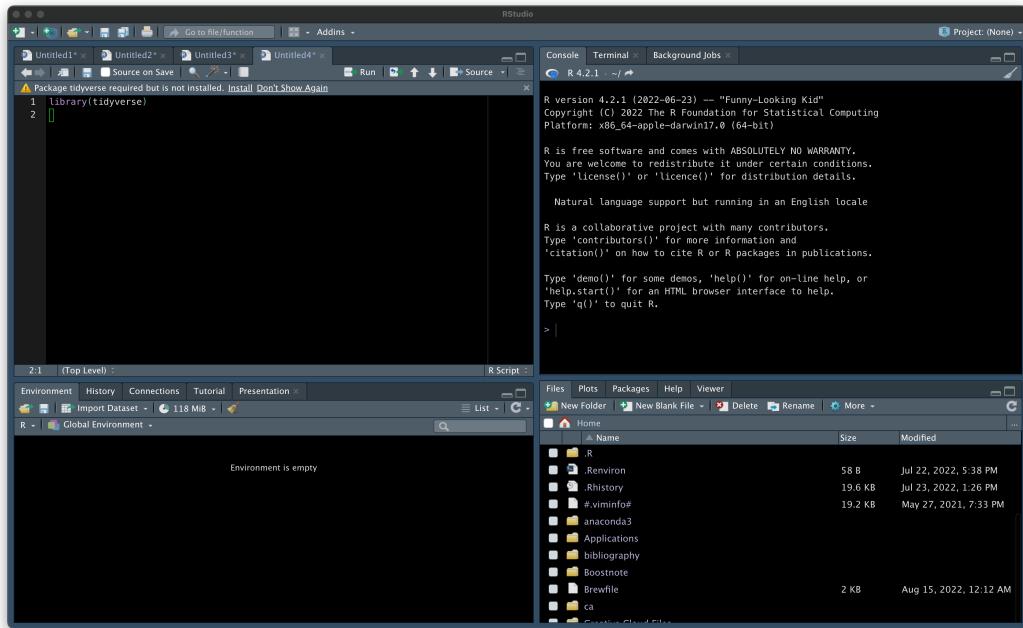


FIGURA 2. Un ambiente de desarrollo, Rstudio.

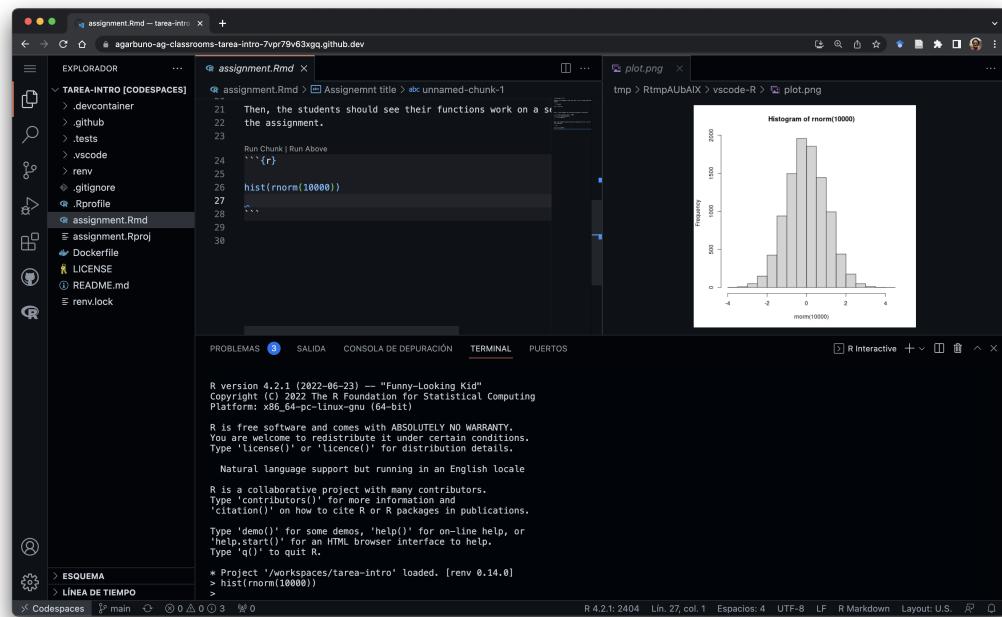
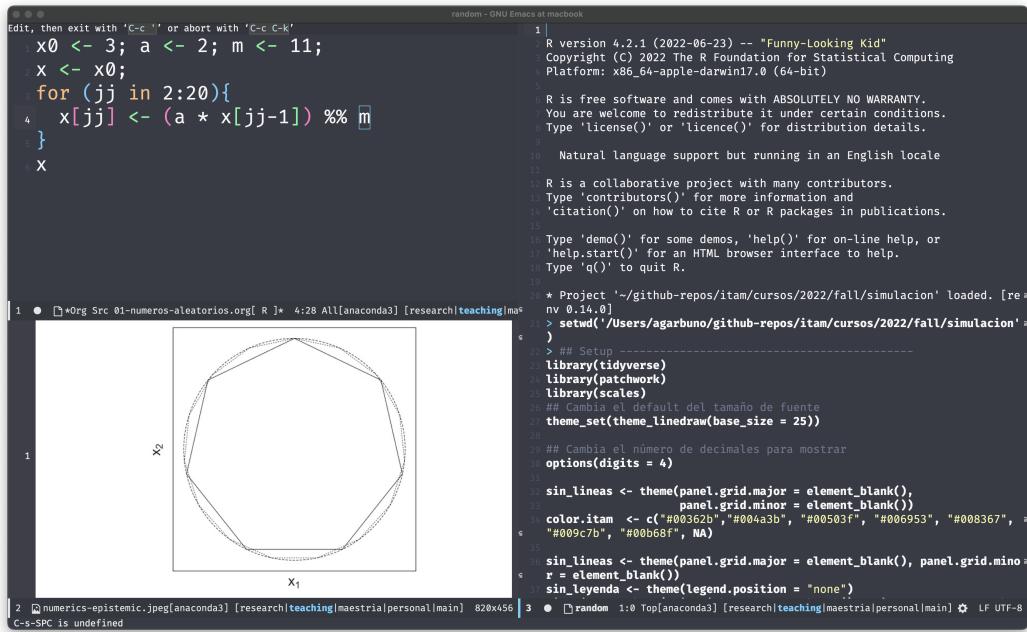


FIGURA 3. Un ambiente de desarrollo general, Visual Code Studio. En la imagen se muestra una sesión en un explorador de internet.



The screenshot shows an Emacs window with two buffers. The top buffer is a terminal window titled 'random - GNU Emacs at macbook' displaying R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid". It contains R code for generating a sequence of points and plotting them. The bottom buffer shows a scatter plot of points forming a circle in a 2D space with axes labeled x1 and x2.

```

Edit, then exit with 'C-c C-j' or abort with 'C-c C-k'
x0 <- 3; a <- 2; m <- 11;
x <- x0;
for (jj in 2:20){
  x[jj] <- (a * x[jj-1]) %% m
}
x

R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

* Project ' ~/github-repos/itam/cursos/2022/fall/simulacion' loaded. [re>
nv 0.14.0]
setwd("~/Users/agarbuno/github-repos/itam/cursos/2022/fall/simulacion")
#> ## Setup
library(tidyverse)
library(patchwork)
library(scales)
## Cambia el default del tamaño de fuente
theme_set(theme_linedraw(base_size = 25))

## Cambia el número de decimales para mostrar
options(digits = 4)

sin_linesas <- theme(panel.grid.major = element_blank(),
                      panel.grid.minor = element_blank())
color_itam <- c("#00362b", "#004a3b", "#00503f", "#006953", "#008367",
                 "#009c7b", "#00b68f", NA)

sin_linesas <- theme(panel.grid.major = element_blank(), panel.grid.mino>
r = element_blank())
sin_leyenda <- theme(legend.position = "none")

numerics-epistemic.jpeg[anaconda3] [research|teaching|maestria|personal|main] 828x456 | 3
random 1:top[anaconda3] [research|teaching|maestria|personal|main] ⚙ LF UTF-8
C-s-SPC is undefined

```

FIGURA 4. Ambiente de desarrollo basado en Emacs.

8.1. ¿Por qué utilizamos el tidyverse?





8.2. ¿Por qué utilizamos *tidymodels*?

La búsqueda CRAN Task View: Machine Learning & Statistical Learning:

abess (core) ahaz arules BART bartMachine BayesTree BDgraph biglasso bmrM Boruta bst C50 caret CORElearn Cubist deepnet DoubleML e1071 (core) earth effects elasticnet evclass evtree frbs gamboostLSS gbm (core) ggRandomForests glmnet glmpath GMMBoost gradDescent grf grplasso grpreg h2o hda hdi hdm ICEbox ipred islasso joineR kernlab (core) klaR lars lasso2 LiblineaR maptree mboost (core) mlpack mlr3 mlr3proba mpath naivebayes ncvreg nnet (core) OneR opusminer pamr party partykit pdp penalized penalizedLDA picasso plotmo quantregForest randomForest (core) randomForestSRC ranger rattle Rborist RcppDL rdetools relaxo rgenoud RGF RLT Rmalschains rminer ROCR RoughSets rpart (core) RPMM RSNNS RWeka RXshrink sda SIS splitTools ssgraph stabs SuperLearner svmpath tensorflow tgp torch tree trtf varSelRF wsrf xgboost

REFERENCIAS

- [1] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6848-6 978-1-4614-6849-3. . [1](#)