Estadística III para In

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Agenda

- anuncios varios
  - Tarea 2 entrega lunes 8 de Mayo 2023(Preguntas
- modelos de analitica (machine learning-ML) Supe
  - Árboles
    - Árboles simples
      - Matemática de los árboles
    - RandomForest
    - GBM

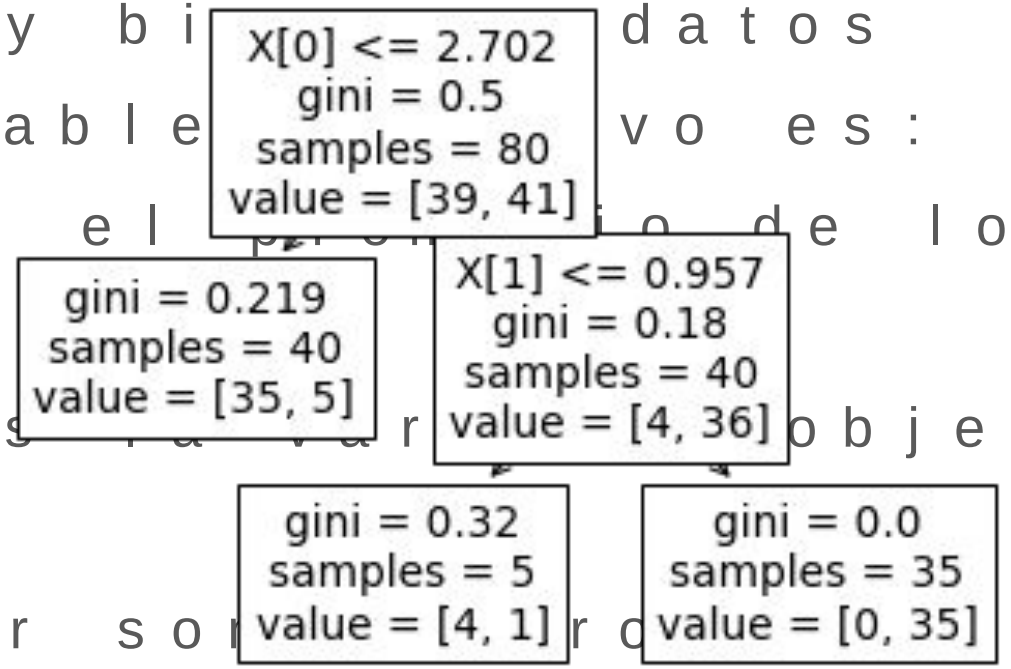
Supervisado, Árboles

Árboles son modelos que separan los datos basados en la capacidad de una variable de dividir los datos. Funcionan muy bien con datos categóricos. Si la variable objetivo es:

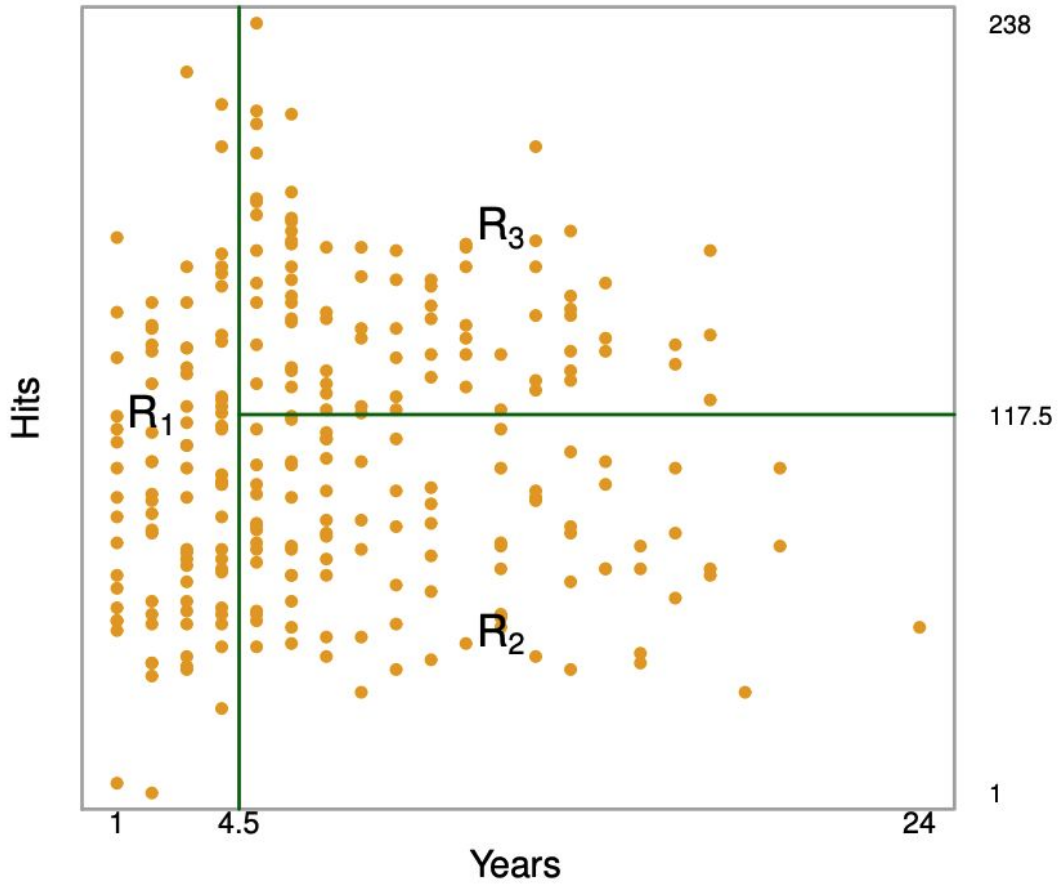
- continúa predicción es el promedio de los vecinos.
- categórica predicción es la variable objetivo más común

Los parámetros a definir son profundidad del árbol, numero mínimo de samples entre otros.

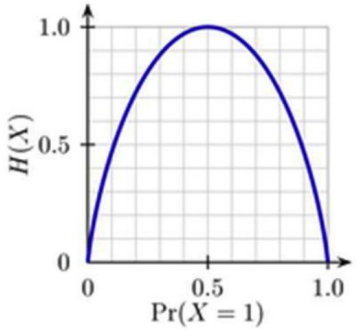
Ejemplo: utilizando datos histór
si un alumno pasará el curso



Entropía , Como funcionan los á



Information Entropy



- For a Bernoulli trial (X = {0,1}) the graph of entropy vs. Pr(X = 1). The highest H(X) = 1 = log(2)

$$H_{(S)} = \sum_{i=1}^{C} -p_i \log_2 p_i$$

$$H_{(T,X)} = \sum_{c \in X} p_{(c)} H_{(c)}$$

$$Gain_{(T,X)} = H_{(T)} - H_{(T,X)}$$

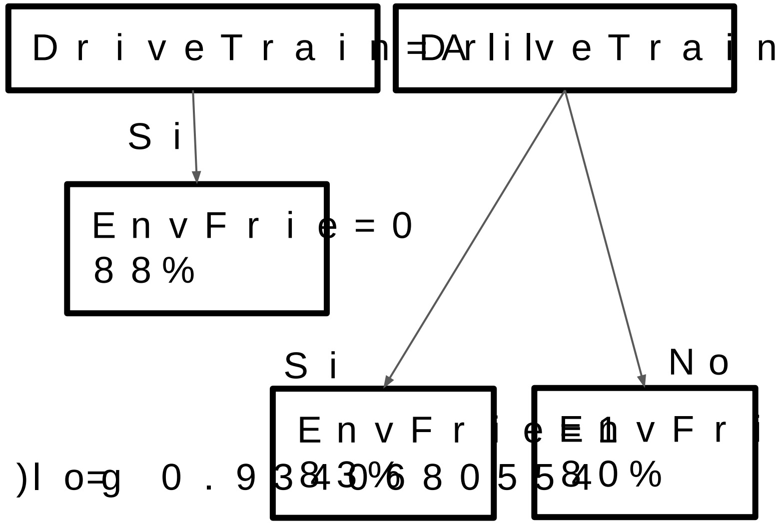
Calcular el árbol a mano para e

Type	DriveTrain	Cylinders	EnvFriendly
Sedan	All	6.0	0
SUV	All	6.0	0
Wagon	Front	4.0	1
Sedan	Rear	8.0	0
Sedan	Front	4.0	1
SUV	Front	4.0	1
SUV	All	6.0	0
Sedan	All	6.0	0
Sedan	All	6.0	0
Sedan	Front	4.0	1
Sports	Rear	6.0	0
Wagon	Rear	6.0	0
Wagon	All	6.0	0
Sedan	Front	6.0	1
Sedan	Rear	8.0	0
SUV	All	6.0	0
Wagon	All	4.0	1
Sports	Rear	4.0	1
SUV	All	8.0	0
Sedan	Front	6.0	0

$$H_{(S)} = \sum_{i=1}^{C} -p_i \log_2 p_i$$

$$H_{(T,X)} = \sum_{c \in X} p_{(c)} H_{(c)}$$

$$Gain_{(T,X)} = H_{(T)} - H_{(T,X)}$$



H(EnvFriendly) = -7/20 $\log_2$(7/20) - 13/20 $\log_2$(13/20) = 0.934068055

H(EnvFriendly, DriveTrain) = p(Train=All)H(Tr
p(Train=Front)H(Train=Front) +
p(Train=Rear)H(Train=Rear) = 0.645689529

Gain(EnvFrindly, DriveTrain) = 0.934068055 - 0.645689529 = 0.2883791025

Gain DriveTrain es mayor que la ganancia de la variable Type

* A Course of Machine Learning http://ciml.info/

Algoritmo árboles simples

Algorithm 8.1 *Building a Regression Tree*

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.

2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.

3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:

   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.

   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results for each value of α, and pick α to minimize the average error.

4. Return the subtree from Step 2 that corresponds to the chosen value of α.

* A Course of Machine Learning http://ciml.info/
* Gareth James, An Introduction to Statistical Learning

Bootstrap y Árboles

"The bootstrap is a widely applicable and extremely powe
the uncertainty associated with a gi
Introduct
Learning

Consiste en crear muchos experimentos tomando pequeñas
reducir la varianza.

* A Course of Machine Learning http://ciml.info/

Bagging y Árboles

"Bootstrap aggregation, or bagging, is a general-purpose learning method; we introduce it here because it is par trees"

Taken from An Introduction to Statistical Learning

Uso: Entrenar muchos árboles y crear un promedio ponde

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

* A Course of Machine Learning http://ciml.info/

Random Forest (Árboles)

"Random forests provide an improvement over bagged
decorrelates the trees. As in bagging, we build a nu
But when building these decision trees, each time a
predictors is chosen as split candidates from the
Learning

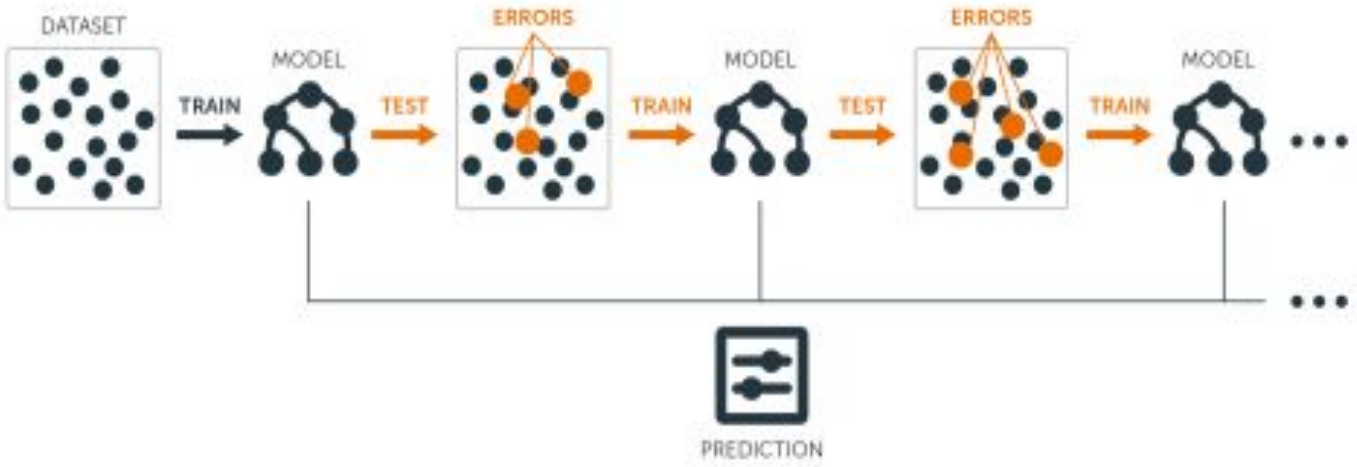
Tiada

An Introduc

Entrenar muchos árboles en paralelo utilizando diferentes muestras

Boosting GBM(Árboles)

"Boosting works in a similar way, except that the t
information from previously grown trees. Boosting do
on a modified version Taken from An Introduction to Statistica



Boosting GBM (Árboles)

Algorithm 8.2 *Boosting for Regression Trees*

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.

2. For $b = 1, 2, \ldots, B$, repeat:

   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).

   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x). \tag{8.10}$$

   (c) Update the residuals,

   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \tag{8.11}$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \tag{8.12}$$