

SESIÓN 5 PROGRAMACIÓN EN R

MINERÍA DE DATOS EN R.

Hugo Andrés Dorado B.



Contenido

- Definiciones en minería de datos.
- Tipo de aprendizaje.
- Algoritmo de predicción.
- Tipos error.
- Sobre parametrización.
- Diseño del estudio.
- Validación cruzada.

Definiciones


- **Big data:** es una tendencia que hace referencia al **almacenamiento de grandes volúmenes de datos** y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos.
- **Minería de datos:** Es un campo de las ciencias de la computación que tiene como propósito descubrir **patrones en grandes** volúmenes de conjuntos de datos.
Utiliza los métodos de la inteligencia artificial, aprendizaje automático y estadística.

Definiciones

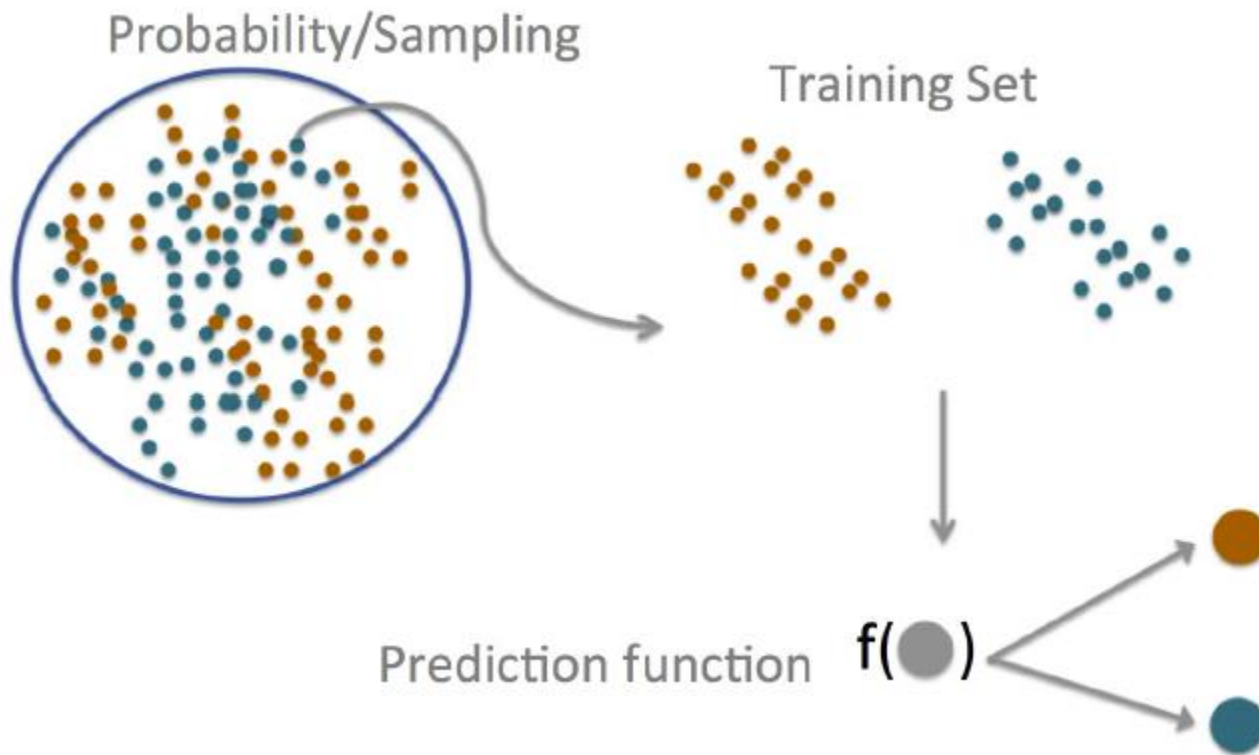
- **Características, variables de entrada (Features):** Variables medidas sobre las observaciones que se asocian luego a un variable salida.
- **Variable de salida:** Variable a explicar de interés.
- **Función costo:** Es una función que permite aproximar un conjunto de variables de entrada para generar una respuesta aproximada según la variable de salida.

$$Y = f(X) + \varepsilon$$

Tipo de aprendizaje

- **Aprendizaje supervisado:** se deduce una función, de acuerdo a un conjunto de variables de salida para la reducción de un error.
 - Clasificación.
 - Regresión.  Predicción o interpretación
- **Aprendizaje no supervisado:** No hay una variable de salida, se busca compresión de los datos tratando un conjunto de variables de entrada.
 - Clustering.
 - Componentes principales.

Predicción en modelos de machine learning



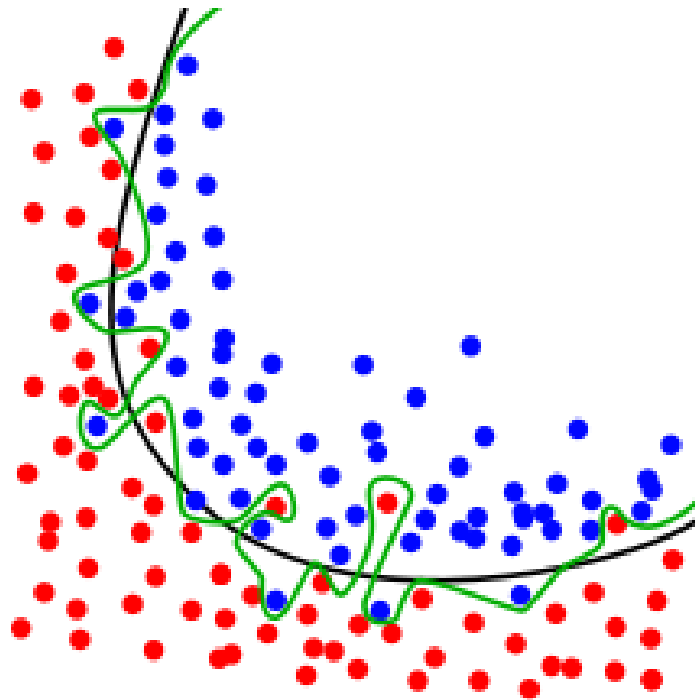
Algoritmos de predicción

- **Pregunta:** ¿Que mails son spam?, ¿Que zonas son bosque?, ¿Que clientes serán morosos?
- **Entrada de datos:** conjuntos de e-mail, Imágenes satelitales, información de clientes. (Ya clasificados)
- **Variables de entrada:** Frecuencia de ciertas palabras, índices espectrales por color, variables seleccionadas
- **Algoritmo:** Redes neuronales artificiales, support vector machine, J46
- **Parámetros:** (Tasa de decaimiento, neuronas ocultas) ,(costo), (umbral de confianza)
- **Evaluación.** (Precisión, exactitud, concordancia)

Tipos de error.

- **Error dentro de la muestra:** La tasa de error que se obtiene en los mismos datos para construir el modelo.
- **Error fuera de la muestra:** La tasa de error que se obtiene al traer nuevos datos no mostrados, también conocido como error de generalización.

Overfitting – sobre parametrización

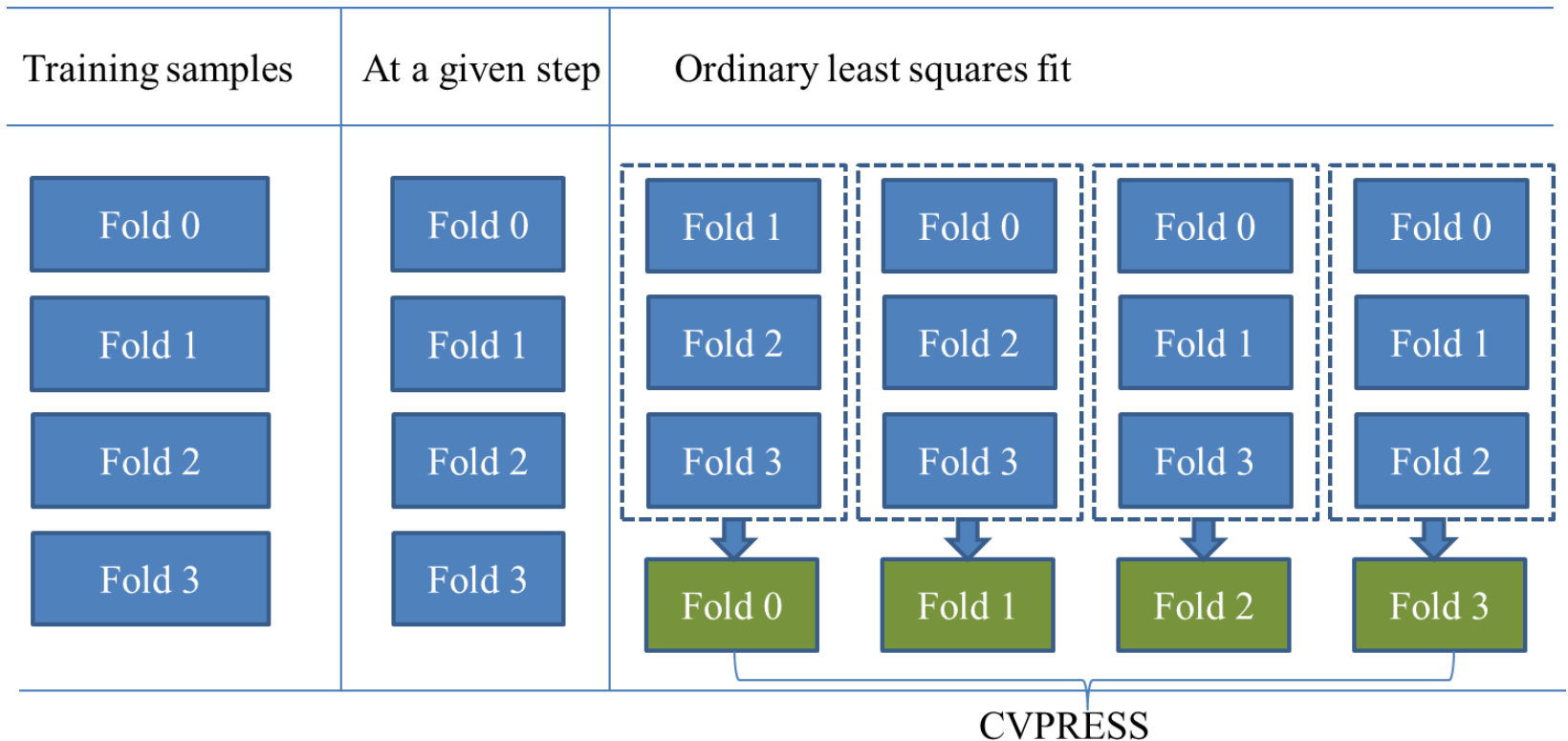


Generalización
Nuevos datos

Diseño de estudio para el conjunto de datos

1. Definir una tasa de error.
2. Partir el conjunto de datos en:
Entrenamiento, prueba y validación (opcional)
(60,20,20) Grandes; (60,40) medianos
3. Sobre el conjunto de entrenamiento hacer selección de variables de entradas
4. Sobre el conjunto de entrenamiento realizar optimización de parámetros. (Utilizar cross validation)
5. Validar de acuerdo a la tasa de error.

K - fold



Medir el desempeño

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

TP: Verdadero positivo.

FP: Falso positivo.

FN: Falso negativo.

TN: Verdadero negativo

Mean squared error (MSE):

$$\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2$$

Root mean squared error (RMSE):

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2}$$

Sensitivity

$$\rightarrow TP / (TP+FN)$$

Specificity

$$\rightarrow TN / (FP+TN)$$

Positive Predictive Value

$$\rightarrow TP / (TP+FP)$$

Negative Predictive Value

$$\rightarrow TN / (FN+TN)$$

Accuracy

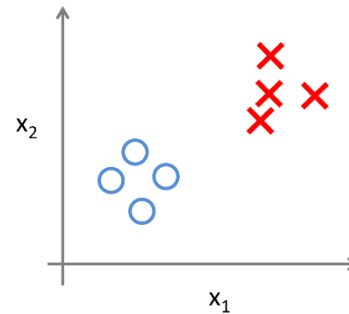
$$\rightarrow (TP+TN) / (TP+FP+FN+TN)$$

Métodos en machine learning para implementar

Modelos supervisados.

- Redes neuronales artificiales
- Árboles de clasificación y regresión.
- Random forest.
- Support vector machine.

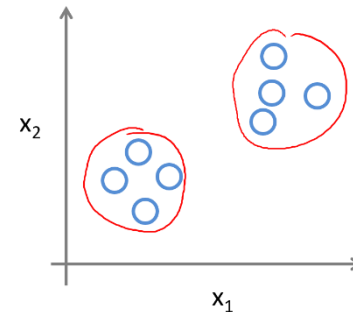
Supervised Learning



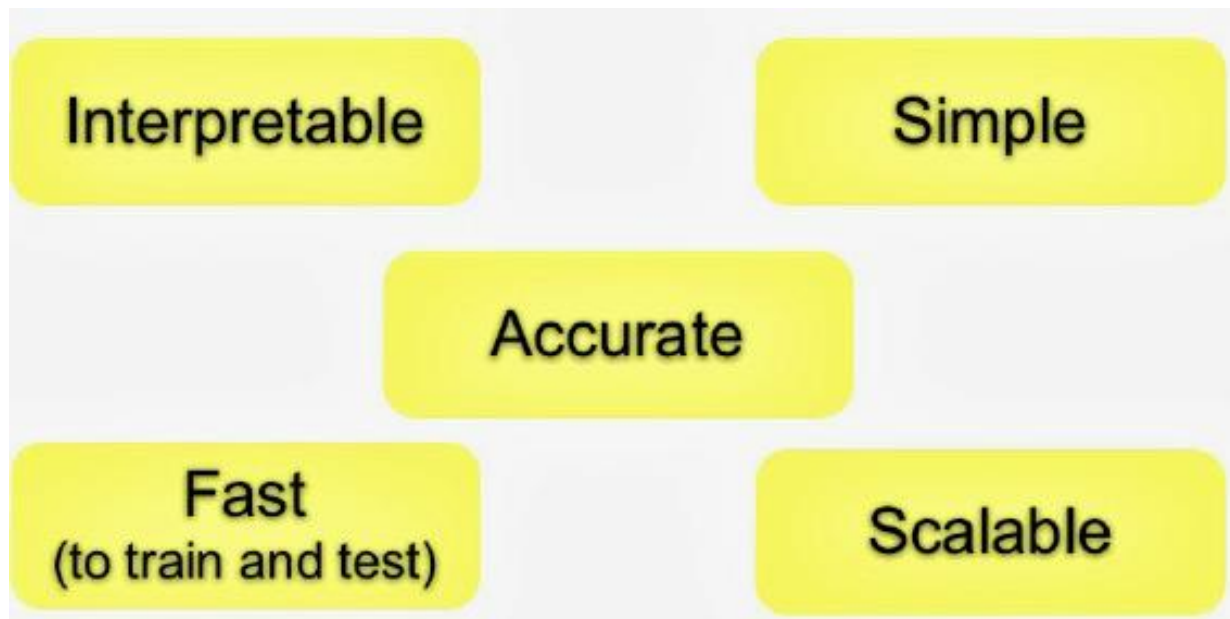
Modelos no supervisados.

- Cluster jerárquico.
- Kmeans
- PCA

Unsupervised Learning



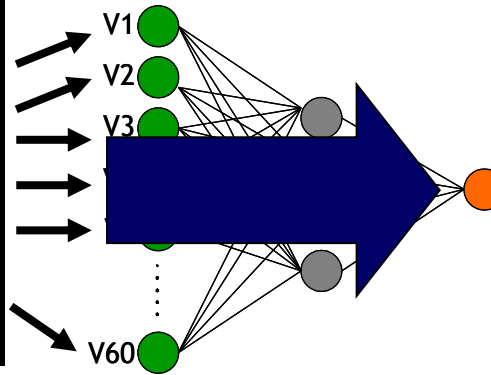
El mejor método de aprendizaje de máquina



Neural networks (Multilayers perceptron)

	V1	V2	V3	V4	V5	...	V60	L 1	L 2	L 3	L 4	L 5	...	Kg/lote
Obs 1	0.1	18	3	312	0.3	...	89	0	1	0	1	0	...	2.39
Obs 2	0.2	15	4	526	0.1	...	52	1	0	0	0	1	...	30.35
Obs 3	0.6	14	1	489	0.2	...	64	0	1	1	1	1	...	42.25
Obs 4	0.05	19	2	523	0.5	...	13	0	0	0	0	1	...	52.50
Obs 5	0.4	13	3	214	0.6	...	57	1	1	1	1	1	...	
Obs 6	0.8	12	4	265	0.4	...	24	1	1	0	1	0	...	82.25
Obs 7	0.2	15	1	236	0.8	...	26	0	0	1	0	0	...	89.28
Obs 8	0.1	17	3	541	0.1	...	35	0	1	1	1	0	...	125.0
Obs9	0.6	16	2	845	0.3	...	51	0	0	1	1	0	...	142.8
Obs10	0.1	18	1	126	0.1	...	43	1	1	0	0	1	...	150.0
...
Obs3000	0.04	15	3	235	0.6	...	85	1	1	1	1	0	...	180

Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
0.1	0.2	0.6	0.05	0.4	0.8	0.2	0.1	0.6	0.1	...	0.04
18	15	14	19	13	12	15	17	16	18	...	15
3	4	1	2	3	4	1	3	2	1	...	3
312	526	489	523	214	265	236	541	845	126	...	235
0.3	0.1	0.2	0.5	0.6	0.4	0.8	0.1	0.3	0.1	...	0.6
...
89	52	64	13	57	24	26	35	51	43	...	85



Predicted

Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
2.07	29.0	53.5	50.5		89.5	99.2	120	172	170	...	188

Observed

Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
2.3	30.3	42.5	52.5		82.2	89.2	125	142	150	...	180

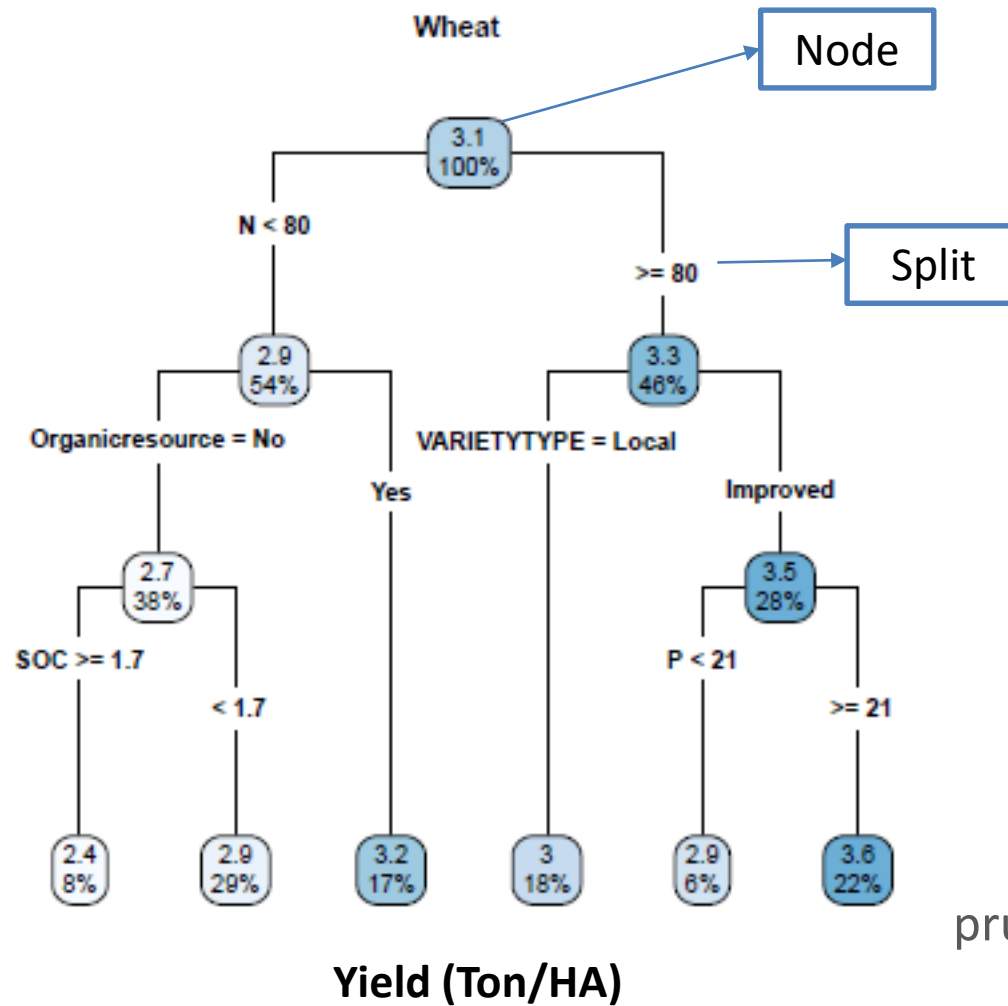


CART(Classification and regression trees)

Index

Gini

information

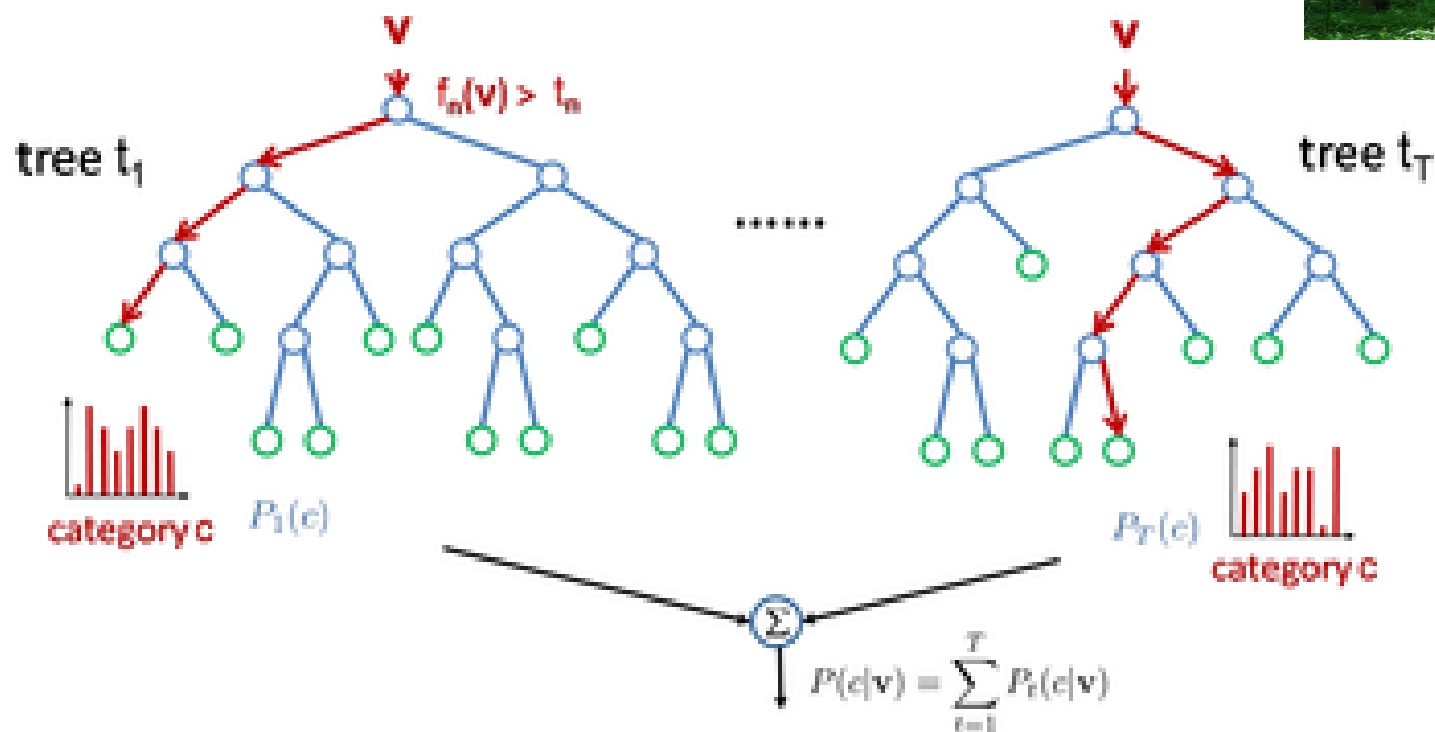


pruning

Random forest

mtry = number of variables

ntrees = number of trees

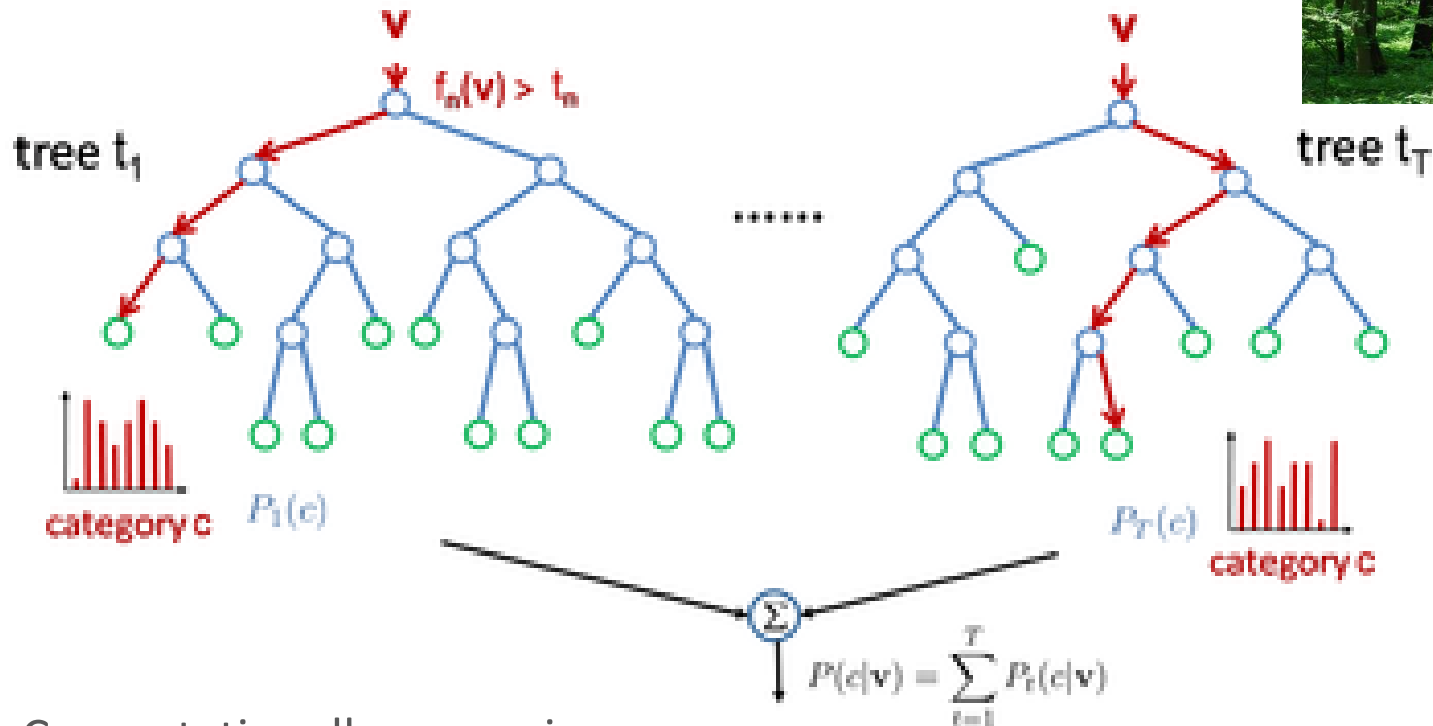


The split is based
in gini coefficient
or information
index

Conditional forest

mtry = number of variables

ntrees = number of trees



The split is based in permutation tests

Computationally expensive
Reduce the random forest bias

Resumen

Buscar datos:

- <https://archive.ics.uci.edu/ml/datasets.html>

Filtrar datos:

- Funciones básicas en R, desde la lectura.

Análisis exploratorio

Transformar datos:

- Merge, dcast, plyr

Determinar estrategia de partición de datos de entrenamiento y validación.

- Seleccionar atributos.
- Optimizar parámetros.

Escoger el modelo final, resultados y pruebas de tasa de error.

Transferir resultados a usuario.

Bibliografía

- <http://caret.r-forge.r-project.org/>
- <http://www.rdatamining.com/>
- <http://ucanalytics.com/blogs/learn-r-12-books-and-online-resources/>
- <https://www.coursera.org/specializations/jhu-data-science>

This is the end ☹️

GRACIAS!!!!

