Machine Learning en las Ciencias Sociales

¿Con qué se come?

Germán Rosati (CONICET-UNSAM / PIMSA)

Datos

MAS_500 Aglomerados segun tamanio	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y anio)
N	8	108	2	2	03/06/1990
N	8	108	3	2	29/12/2005
N	8	108	3	1	26/01/2018
N	8	108	1	2	30/03/1978
N	8	108	3	2	20/09/2009
N	8	141	1	1	26/04/1967
N	8	221	1	1	15/03/1955
N	8	221	2	2	25/04/1956
N	8	221	3	2	10/06/1994
N	8	221	1	1	22/07/1944
N	8	221	3	1	23/08/1985
N	8	309	1	1	14/06/1976
N	8	309	2	2	17/06/1978
N	8	309	3	2	20/07/1997
N	8	309	3	1	19/10/2001
N	8	309	1	2	02/01/1967
N	8	309	3	2	29/06/1982
N	8	88	1	1	15/08/1974

14/06/1976

Datos

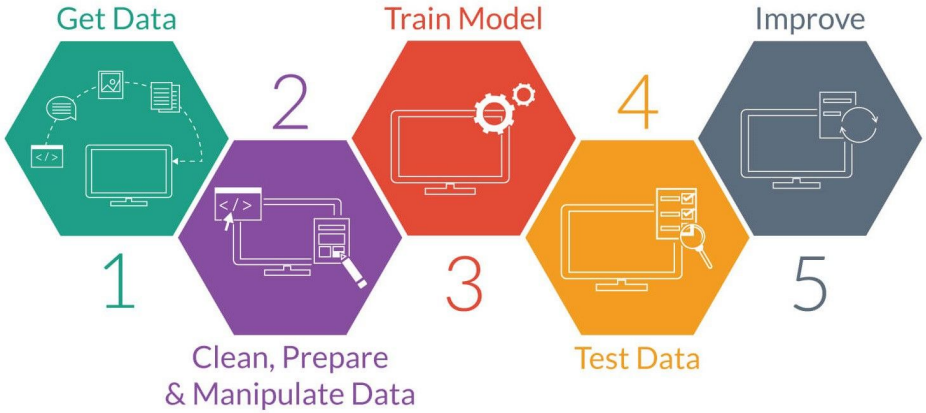
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 3

[1] a bailar a bailar | que la orquesta se va | sobre el fino garabato | de un tango nervioso y lerdo | se ira borrando el recuerdo | a bailar a bailar | que la orquesta se va | el ultimo tango perfuma la noche | un tango dulce que dice adios | la frase callada se asoma a los labios | y canta el tango la despedida! | vamos! a bailar! | tal vez no vuelvas a verla nunca | y el ultimo tango perfuma la noche | y este es el tango que dice el adios | a bailar a bailar | que la orquesta se va! | quedara el salon vacio | con un monton de esperanzas | que iran camino al olvido | a bailar a bailar | que la orquesta se va!

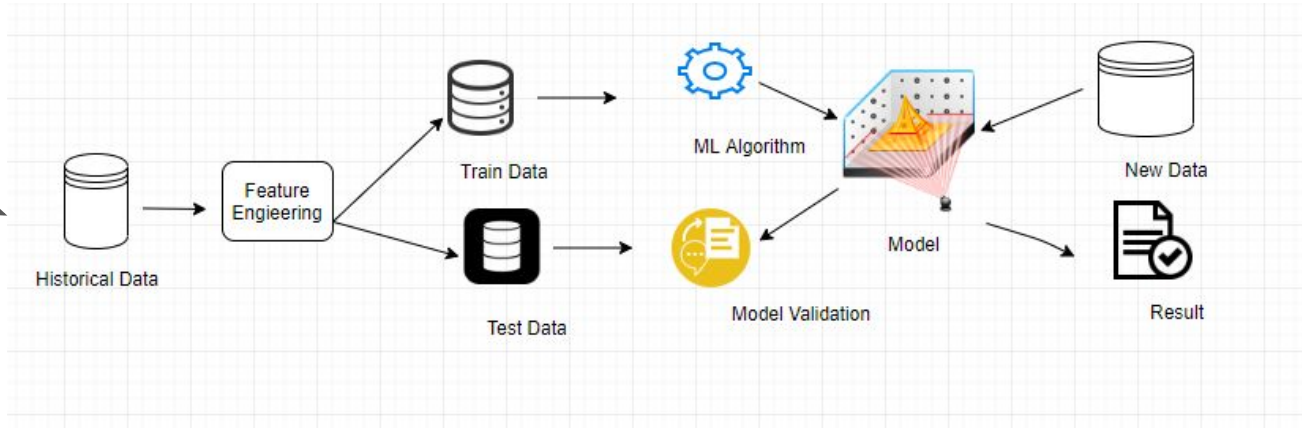
[2] este tango nacio para bailarse | y asi hamacarse muy suavemente | oigan ustedes este compas… | es muy sencillo bailar el tango | un doble paso despues descanso | la media vuelta la vuelta entera | y siempre junto a la compañera | este tango nacio para bailarse | no hay que quedarse mirandolo

[3] nacio en la calle quito | entre boedo y colombres | barrio de tauras de hombres | de timbas y de garitos | mi recuerdo es muy estricto | de proscenio un corralon | modesto fue su blason | y la dulce purretita | se lavaba la carita | en el viejo pileton | amante del variete | soñaba con ser artista | comenzo como corista | hasta llegar a vedette | piernas tipo mistinguette | cintura bien contorneada | anatomia envidiada | y un rostro angelical | para que plumas y percal | lucieran como hermanadas | siempre causo sensacion | en cine radio y teatro; | se volco al dos por cuatro | con sentida emocion | triunfo en television | y nadie podra dudar | fue figura consular | en todos los escenarios | recogio aplausos a diario | se llamaba beba bidart

El problema… o volviendo a Método I



Problema, pregunta de investigación



Algoritmos y preguntas de investigación

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

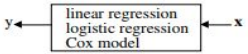
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

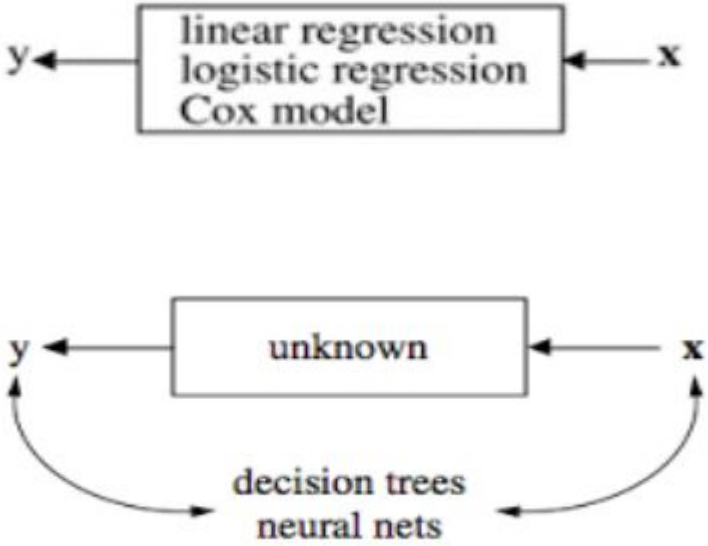
1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.



Ciencias Sociales y Machine Learning

- Como “auditores” de los modelos (de esto va a hablar Vicky)

Ciencias Sociales y Machine Learning

- Como “auditores” de los modelos (de esto va a hablar Vicky)

- **Como usuarios o consumidores (de esto van a hablar mucho Tomás, Natsu y Tonio)**

Ciencias Sociales como usuarias de ML

Automatización de procesos para la construcción de bases de datos de protestas

[Hanna, 2017]

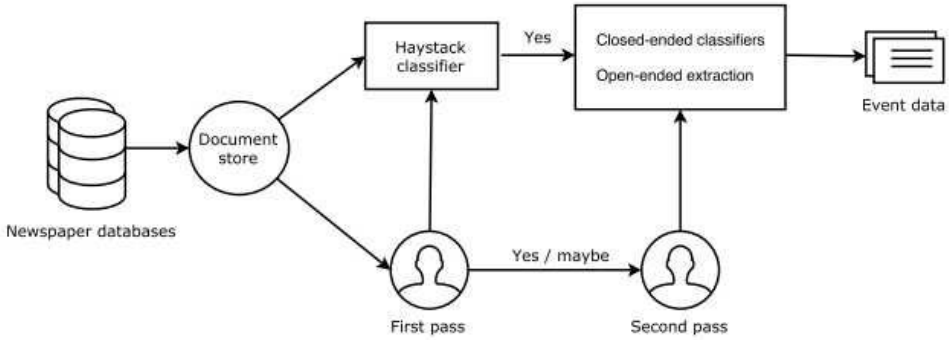


Figure 1: MPEDS pipeline with training.

Ciencias Sociales como usuarias de ML

Integración de comunidades inmigrantes en grandes ciudades

[Lamanna,Lenormand, et al 2016]

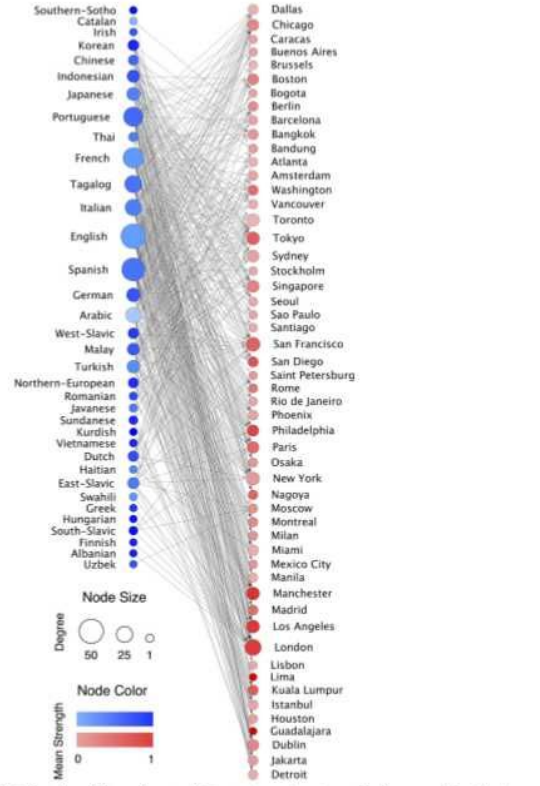


Fig 2. Bipartite spatial integration network. The network comprises of two sets: L of Languages and C of cities; the languages detected are connected to the cities set where the corresponding community of immigrants has been found. The weight of the edge corresponds to the values of h_{lc}. The size of the nodes is proportional to its degree and the color to its mean strength.

Ciencias Sociales y Machine Learning

- Como “auditores” de los modelos (de esto va a hablar Vicky)

- Como usuarios o consumidores (de esto van a hablar mucho Tomás, Natsu y Tonio)

- **Como productores (de nuevos métodos y problemas)**

¿Qué leer?


+ práctica
- práctica
- complejidad
+ complejidad
O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
Jake VanderPlas
O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
Aurélien Géron
DEEP LEARNING with Python
François Chollet
Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
big data
walter sosa escudero
breve manual para conocer la ciencia de datos que ya invadió nuestras vidas
MACHINE LEARNING
TOM M. MITCHELL
DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville
Machine Learning
A Probabilistic Perspective
Kevin P. Murphy
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE STATISTICAL INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE

FACTOR DATA

BIG DATA Y CIENCIAS SOCIALES

Investigación básica

Investigación aplicada

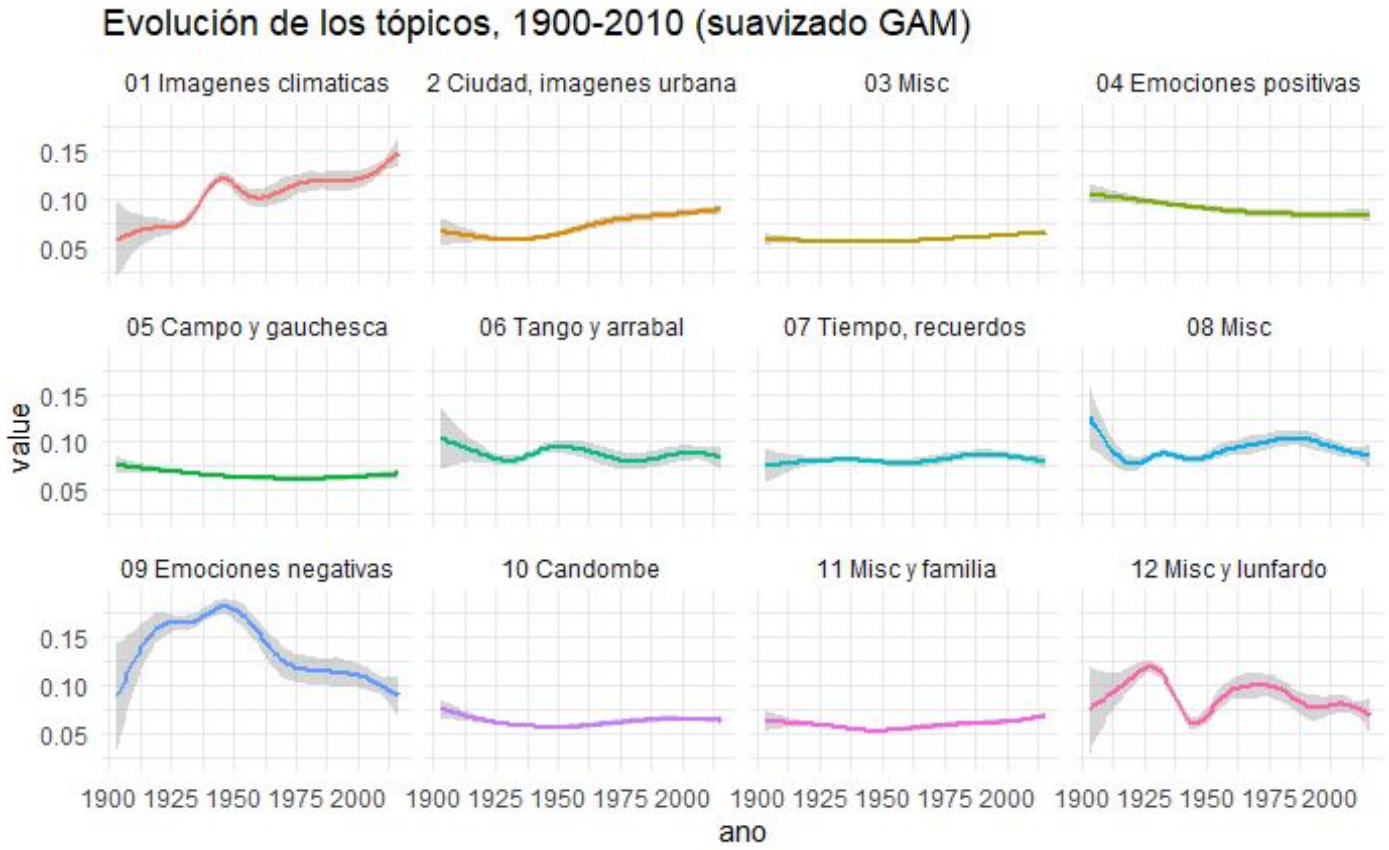
Formación y pedagogía

UNIVERSIDAD NACIONAL DE SAN MARTÍN

IDAES

FACTOR·IDAES
experimentar en ciencias sociales

Algunos proyectos: ¿de qué habla el tango?



¿Preguntas?

@Crst_C

german.rosati@gmail.com

https://gefero.github.io/