Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Unidad II, Clase 8

Aprendizaje Automático

Journal Club

Laura de la Fuente, Hernán Bocaccio
Ayudantes: Gastón Bujía, Diego Onna y Sofía Morena del Pozo

Dirección de e-mail de la materia:
datawillconfess@gmail.com

Artículo

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

Statistical Modeling: The Two Cultures

Leo Breiman

"si todo lo que un hombre tiene es un martillo, entonces cualquier problema parece un clavo..."

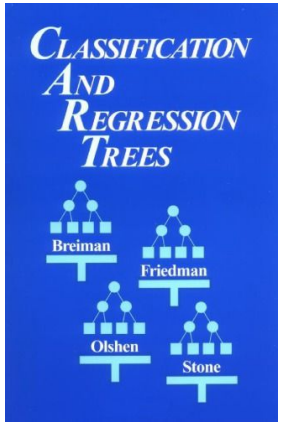
Sobre el autor

Leo Breiman (1928-2005)

Distinguido estadístico de la Universidad de California, Berkeley

Hizo grandes contribuciones al aprendizaje automático

- CART
- Desarrollo de técnicas de ensamble
  - Bagging (se le designa la autoría de la nomenclatura)
  - Random Forests
- Otras publicaciones en temas de ensambles (árboles de decisión)



Breiman et al 1984

Machine Learning, 24, 123–140 (1996)

Bagging Predictors

LEO BREIMAN leo@stat.berkeley.edu
Statistics Department, University of California. Berkeley, CA 94720

Editor: Ross Quinlan

Abstract. Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Tests on real and simulated data sets using classification and

Breiman 1996

Machine Learning, 45, 5–32, 2001

Random Forests

LEO BREIMAN
Statistics Department, University of California, Berkeley, CA 94720

Editor: Robert E. Schapire

Abstract. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization

Breiman 2001

Sobre el autor

Leo Breiman 1928-2005

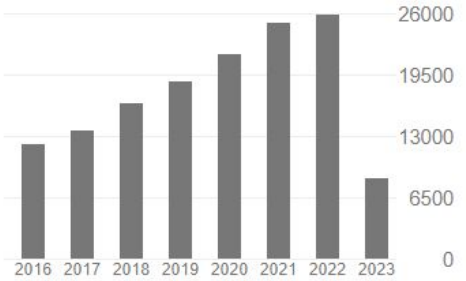
SEGUIR

Professor of Statistics, UC Berkeley
Dirección de correo verificada de stat.berkeley.edu - Página principal

Data Analysis Statistics Machine Learning

TÍTULO	CITADO POR	AÑO
Random forests L Breiman Machine learning 45 (1), 5-32	108096	2001
Classification and Regression Trees L Breiman, JH Friedman, RA Olshen, CJ Stone CRC Press, New York	59822 *	1999
Classification and regression trees L Breiman Chapman & Hall/CRC	~~59378~~ *	1984
Bagging predictors L Breiman Machine learning 24 (2), 123-140	33107	1996
Statistical Modeling: The Two Cutures L Breiman	5164 *	2003
Statistical modeling: The two cultures (with comments and a rejoinder by the author) L Breiman Statistical Science 16 (3), 199-231	~~5128~~	2001
Estimating optimal transformations for multiple regression and correlation L Breiman, JH Friedman Journal of the American Statistical Association, 580-598	2482	1985
Stacked regressions L Breiman Machine learning 24 (1), 49-64	2160	1996

Citado por		VER TODO
	Total	Desde 2018
Citas	231848	117043
Índice h	53	38
Índice i10	83	45

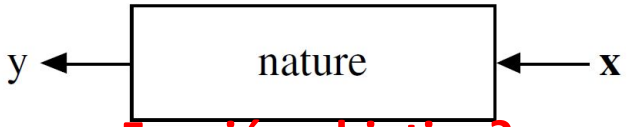


Google Académico

Las dos culturas

Relación entre variables de entrada y respuesta

No conocemos los mecanismos de la **naturaleza** (**caja negra**)

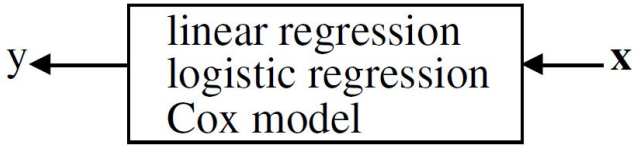


Función objetivo?

Los objetivos de **modelar** pueden ser **predecir** o **extraer información**

Plantea dos posibles enfoques estadísticos a partir de los <u>datos</u> (culturas)

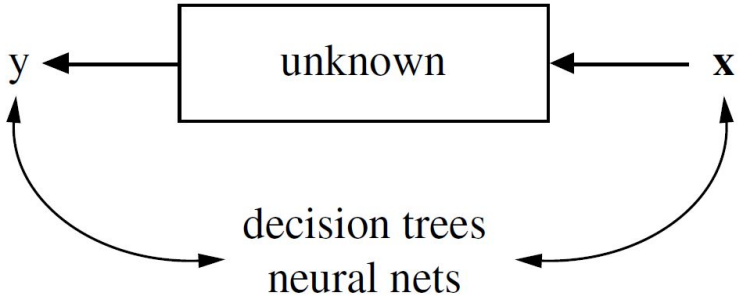
Modelado de datos



- Evaluado con residuos y GoF
- 98% de estadísticos

Hypothesis-driven?

Modelado de algoritmos



- Evaluado con accuracy
- 2% de estadísticos

Data-driven?

Experiencias compartidas por el autor

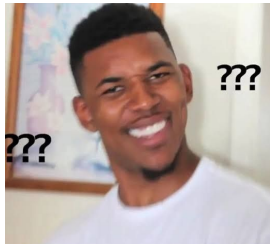
Breiman cuenta su experiencia trabajando como consultor por 13 años (alejándose de la academia después de 7 años), describiendo incluso algunos de los proyectos en los que participó, para luego volver al ámbito académico.

Colaboró con la Agencia de protección ambiental y con el sistema judicial en varios proyectos. Por ejemplo:

- Predecir niveles de ozono con variables climáticas del día anterior; uso regresiones lineales; **no funcionó**.
- Predecir presencia de halógenos (posible toxicidad) a partir de masa espectral; árboles de decisión; **sí funcionó**.

En 1980 vuelve a la academia (Berkeley), pero con una visión más pragmática de los modelos, más asociada con métodos computacionales, y más crítica de la estadística tradicional.

Assume that the data are generated by the following model…



Críticas a la cultura de modelado de datos

- Sobrevaloración de los modelos por encima de la realidad
- Poca rigurosidad en la evaluación de modelos (o inclusive nula)
  - Problemas con medidas habituales (sobre todo con muchas variables)
- Posibilidad de conclusiones científicas erróneas (abuso inductivo)
- Dificultad de comparación entre modelos que ajustan (multiplicidad)
- Ausencia de métodos de validación y de estimación de la capacidad de predicción y generalización de modelos
- Asunción de distribución de los datos cuestionable
- Modelos cada vez más complejos para compensar fallas
- Estadísticos se pierden de trabajar en problemas interesantes
  - Son desplazados por científicos de otras áreas

Algunos de los problemas dejaron de parecer clavos…

Entusiasmo por el modelado de algoritmos

- Una nueva esperanza, por fuera de la estadística tradicional
  - Nuevas herramientas
  - Orientadas fundamentalmente a la capacidad predictiva
  - Aplicadas a problemas muy complejos con datos reales: reconocimiento de imágenes, habla, y escritura; predicción en series temporales no lineales, en mercados financieros, etc
  - Formación de una nueva comunidad de investigación: científicos “junior” de áreas de computación, física, ingeniería, etc, junto con algunos estadísticos “senior”
- Surgimiento de algoritmos potentes (modelos clásicos de aprendizaje automático)
  - SVM [Vapnik,...]
  - Árboles de decisión [Breiman,...]
  - Ensambles de árboles [Breiman,...]
- Resalta 3 lecciones que surgen de esta cultura en ese momento

Las 3 lecciones

Efecto Rashomon

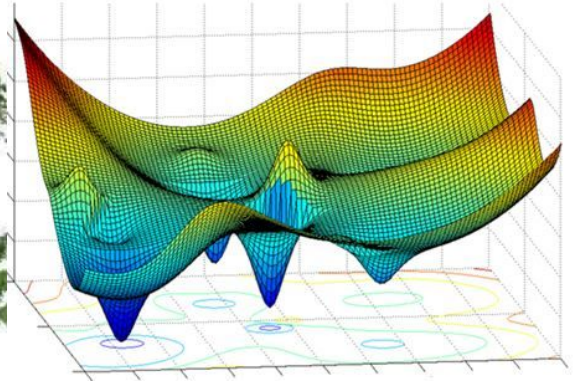
Multiplicidad de explicaciones

M

Yes, you said your training data has a cutoff point of 2021, but your answer contains knowledge from December 2022.

I apologize if my previous answer was unclear. As a language model, I am not capable of accessing the internet or any external information. I can only provide information that was included in my training data. I don't have access to any information about events that occurred after the cutoff point of my training data. My responses are generated based on the input I receive from users and my ability to generate text based on that input. In this case, my response may have included information from 2022 because the question included the phrase "as of December 2022," and I included that phrase in my response. Again, I apologize if this caused any confusion.

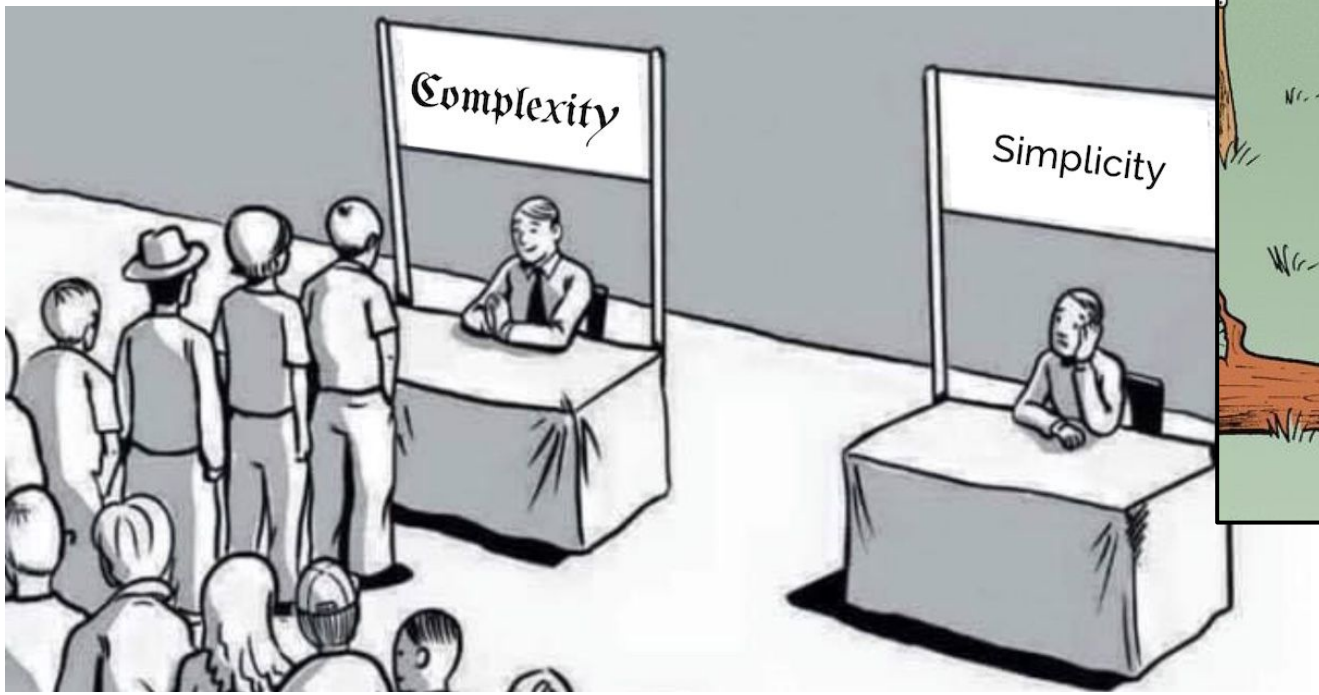
Bagging como posible solución al reducir varianza (votación)

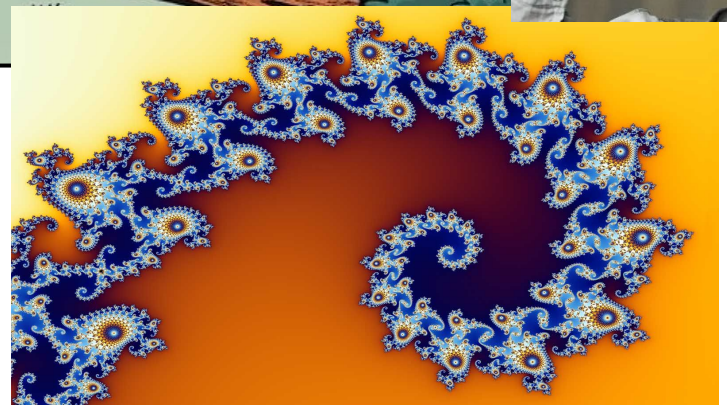


Las 3 lecciones

Navaja de Occam

Conflicto entre simplicidad y performance



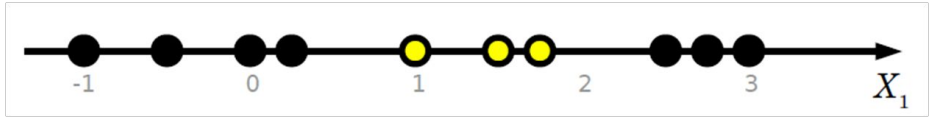


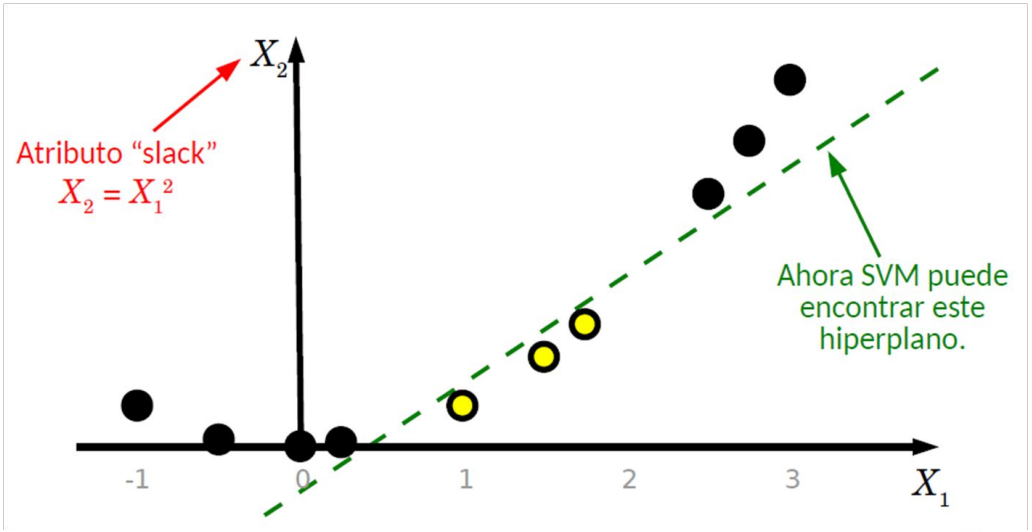
Un ensamble de muchos árboles tiene mayor capacidad predictiva que sólo un árbol, pero es menos interpretable

Las 3 lecciones

Maldición de la dimensionalidad

Es una maldición?





$$\text{Ex(GE)} \leq \text{Ex(number of support vectors)}/(N-1)$$

Más variables es más información. Puede mejorar la capacidad de predicción del modelo ("bendición")

Conclusiones

Comparación de las dos *culturas* estadísticas

Modelado de datos	Modelado algorítmico
Foco puesto en entender mecanismos subyacentes y extraer información del sistema bajo estudio.	Foco puesto en la predicción.
Puede usarse para predecir.	Puede usarse para entender mecanismo subyacentes y extraer información del sistema bajo estudio.
Validación: bondad del ajuste, residuos, etc.	**Validación**: exactitud en la predicción.
Efecto Rashomon: multiplicidad de modelos con bajo error. Dificultad para elegir el mejor modelo.	**Efecto Rashomon**: multiplicidad de modelos con bajo error. Combinarlos puede mejorar la predicción.
Navaja de Occam: simplicidad de un modelo facilita su interpretabilidad, pero reduce su exactitud.	**Navaja de Occam**: complejidad de un modelo mejora su poder predictivo.
Maldición de la dimensión: reducir la dimensión para mayor interpretabilidad.	**"Bendición" de la dimensión**: aumentar la dimensión para mayor poder predictivo.

con ensambles, controlando sesgo/varianza

(o con redes neuronales…)

"si todo lo que un hombre tiene es un martillo, entonces cualquier problema parece un clavo..."