

Métodos de análisis cuantitativos multivariados

Clase 1. Introducción al modelado y regresión lineal simple



Un poco de epistemología... ¿qué es un modelo?

Del rigor en la ciencia

by Jorge Luis Borges (March 1946)

. . . En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el Mapa de una sola Provincia ocupaba toda una Ciudad, y el Mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el Tamaño del Imperio y coincidía puntualmente con él. Menos Adictas al Estudio de la Cartografía, las Generaciones Siguientes entendieron que ese dilatado Mapa era Inútil y no sin Impiedad lo entregaron a las Inclemencias del Sol y los Inviernos. En los Desiertos del Oeste perduran despedazadas Ruinas del Mapa, habitadas por Animales y por Mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas.

— Suárez Miranda: *Viajes de varones prudentes*, libro cuarto, cap. XLV, Lérida, 1658.



Un poco de epistemología... ¿qué es un modelo?

- Básicamente: una manera de proponer hipótesis sobre la forma en que se combinan variables
- En general vamos a estar tratando de general modelos de la forma

$$Y = f(X) + e$$

Diagram illustrating the components of the model equation $Y = f(X) + e$:

- Variable dependiente** (Dependent variable): Y
- El modelo** (The model): $f(X)$
- Variable(s) independiente(s)** (Independent variable(s)): X
- El error, lo que no sabemos** (The error, what we don't know): e

- Todo el problema es estimar $f(X)$, es decir, de qué forma(s) se combinan las X para generar un output
- Una posibilidad es suponer que Y es una combinación lineal de las X



¿Para qué modelar?

Para **predecir**

Tenemos valores de un conjunto de variables independientes (X_1 , X_2 , etc.) y queremos un modelo que **prediga** el valor de la variable dependiente (Y)

Para **inferir / comprender**

Queremos **entender** la relación entre la variable dependiente (Y) y el conjunto de variables independientes (X_1 , X_2 , etc.).



Predicción

- f puede funcionar como una caja negra.
- El error de nuestras predicciones se puede descomponer en dos partes:
 - Reducible: mejorar f
 - Irreducible: ϵ (variables que no incluimos o no podemos medir)



Inferencia

- **f** NO puede funcionar como una caja negra.
- La calidad de nuestros resultados depende de una serie de supuestos acerca de la distribución de los datos.
- Nuestro interés principal son preguntas como:
 - ¿Qué variables X están relacionadas con **Y**?
 - ¿Qué dirección tienen estas relaciones?
 - ¿Qué efecto tiene cada variable **X** en la variable **Y**?
 - ¿Las relaciones son lineales o no lineales?



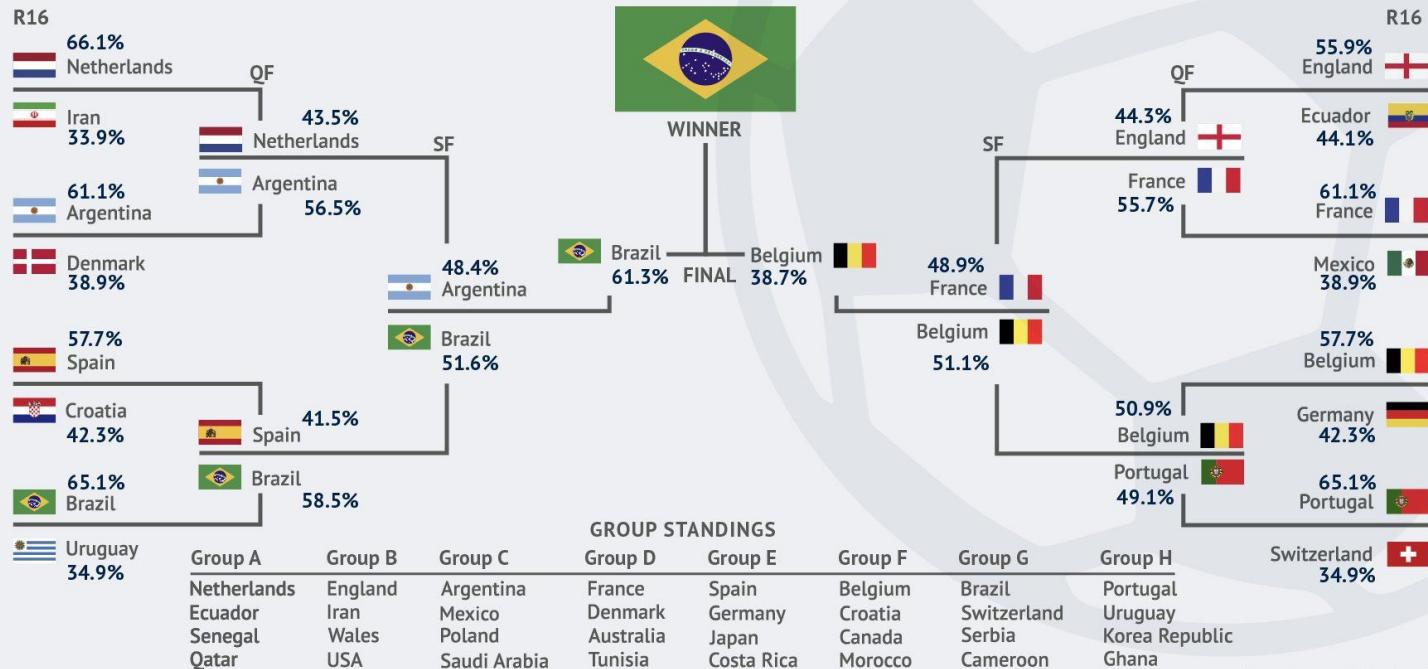
¿Cuándo usamos cada tipo de modelo?

- ¿Qué pasa si nuestro modelo de inferencia no predice bien?
- Problemas para predecir por fuera del intervalo de entrenamiento (e.g.: predecir el futuro con datos del pasado)



Oxford mathematical model predicts route to the 2022 World Cup

@JoshuaABull



% = chance of winning the game



factor-data
EIDAES_UNSAM

Tipos de problemas (Y)

- Aprendizaje Supervisado
 - Variable dependiente => Y, resultado, target
 - Matriz de predictores (p), X, features, variables independientes, etc.
 - Tenemos datos de entrenamiento (conjuntos de X_i, Y_i), observaciones
 - Podemos definir una (o varias) métricas para evaluar los modelos
- Aprendizaje no Supervisado
 - No hay variable target (Y)
 - Solo hay X
 - Es más difícil evaluar qué tan bien funciona el modelo



Tipo de variable dependiente = Y

- Regresión => Y es una variable cuantitativa. El modelo nos proporciona una función que nos da valores predichos de Y para los diferentes valores de X_1, X_2 , etc.
- Clasificación: Y es una variable categórica y el modelo nos da la probabilidad de que cada caso quede en cada categoría.



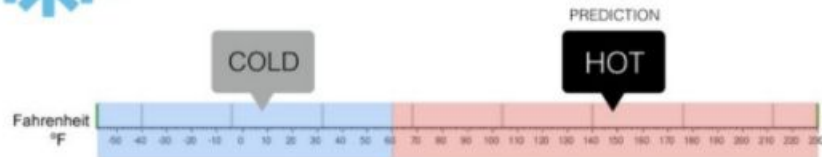
Regression

What is the temperature going to be tomorrow?



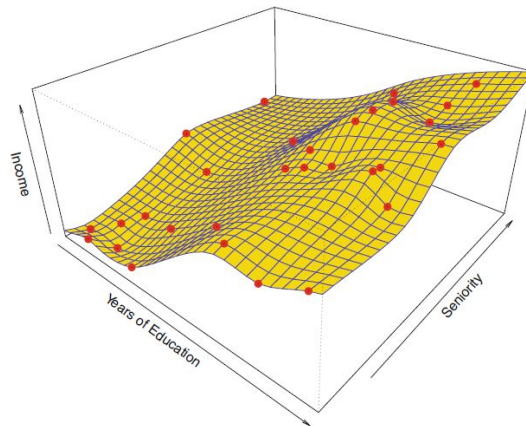
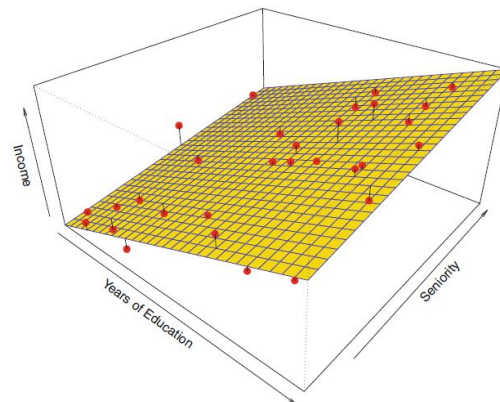
Classification

Will it be Cold or Hot tomorrow?



¿Asumimos una forma del modelo? -f(.)-

- Los modelos paramétricos asumen la forma de la función.
- Los métodos no paramétricos, no, y pueden tener diferentes formas para diferentes valores de X.



Modelos paramétricos

- Suponemos que f tiene cierta forma funcional.
- Ajustamos el modelo.
- Ventajas
 - Suponer una **forma funcional** simplifica la estimación de los parámetros.
- Desventajas
 - La **forma funcional** que elegimos difícilmente coincida con la forma real de f .

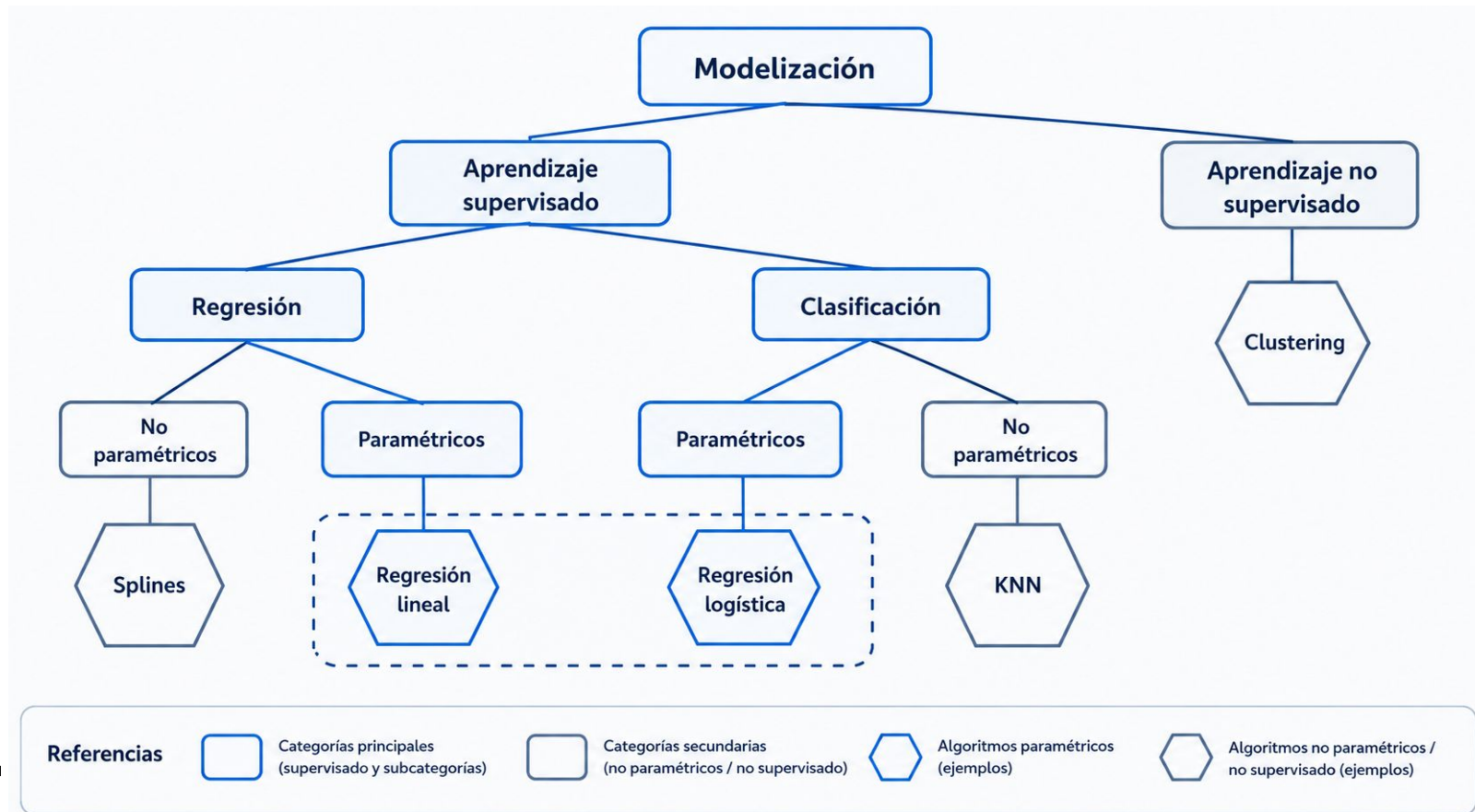


Modelos no paramétricos

- Ventajas
 - Al no suponer una **forma funcional** para **f**, tienen el potencial de ajustarse con precisión a una gama más amplia de formas posibles para **f**.
- Desventajas
 - Necesitamos una cantidad de observaciones mayor que en el caso de los modelos paramétricos.
 - Peligro de overfitting.



Modelos no paramétricos



Vamos al Notebook



factor-data
EIDAES_UNSAM

Geocentric Model



Geocentric Model



Geocentric Model

