Estadística (1937)
Unidad 7.
Introducción a la ciencia de datos

Marcelo Ruiz

Departamento de Matemática, Universidad Nacional de Río Cuarto

2023

Aprendizaje Estadístico

Dato

Todo "dato":

- es una construcción.
- surge como respuesta a preguntas en el contexto de un problema [1].

[1] Imagen tomada de https://twitter.com/Yyomepregunto2/photo

El problema determina el dato

Afirma Chalmers[2] que:

se consiguen hechos relevantes midiendo la concentración de ozono en varios lugares, mientras que no se logra nada midiendo la longitud de los cabellos de 105 jóvenes de Sidney.

[2] Chalmers, A. (1982). *¿Qué es esa cosa llamada ciencia?. Siglo XXI.*

Provisoriamente

¿De qué hablamos cuando hablamos de ciencia de datos?
Hastie, en 2015, sostiene: [3]

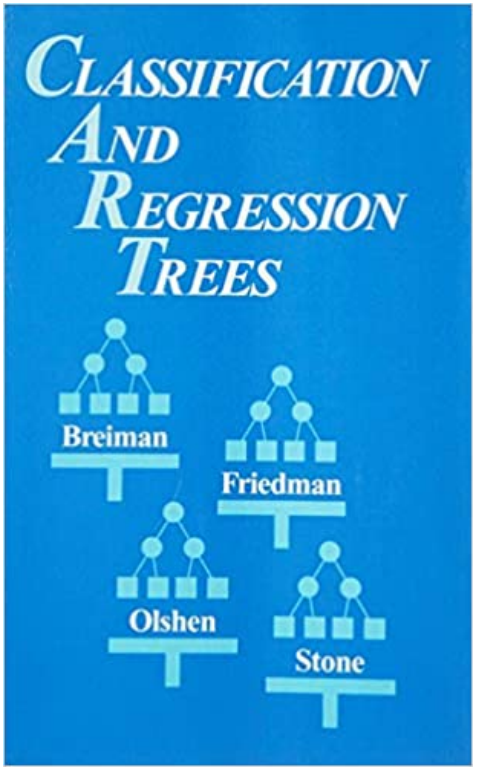
- El Aprendizaje Automático (Machine Learning) construye algoritmos que pueden aprender de los datos.
- El Aprendizaje Estadístico (Satistical Learning) es una rama de la estadística aplicada que emergió en respuesta al AA, enfatizando en los modelos estadísticos y en la evaluación de la incertidumbre.
- La Ciencia de Datos (CD) es la extracción de conocimiento de los datos, utilizando ideas de matemática, estadística, AA, ingeniería,...

[3] https://web.stanford.edu/~hastie/TALKS/SLBD_new.pdf

Las transformaciones científico-técnica y nuevas dinámicas mundiales

Me interesa mencionar algunos aspectos de su emergencia, mucho tiempo atrás.

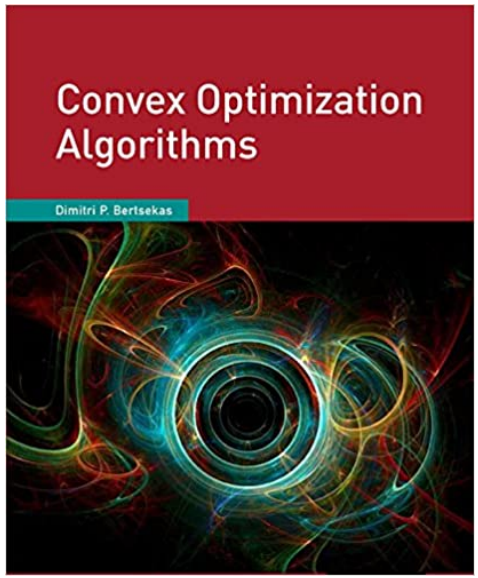
- Ciertas mutaciones en dinámicas disciplinares clásicas y novedosas emergencias:
  - en líneas de la computación y de la estadística.
  - en lingüística (de Peirce a IBM).
  - en la ingeniería de Pattern Recognition.
- Problemas
  - de la industria
  - en inteligencia y defensa (DARPA).



La dinámica de la financiarización y la conectividad

Más recientemente:

- En términos disciplinares, las múltiples articulaciones entre
  - estadística
  - teoría de aproximación.
  - programación convexa
  - ecuaciones diferenciales
  - álgebra lineal numérica
  - optimización
  - teoría de probabilidad
  - ingeniería de software, etc.
- Las fuerzas fácticas de las dinámicas financieras y la conectividad.



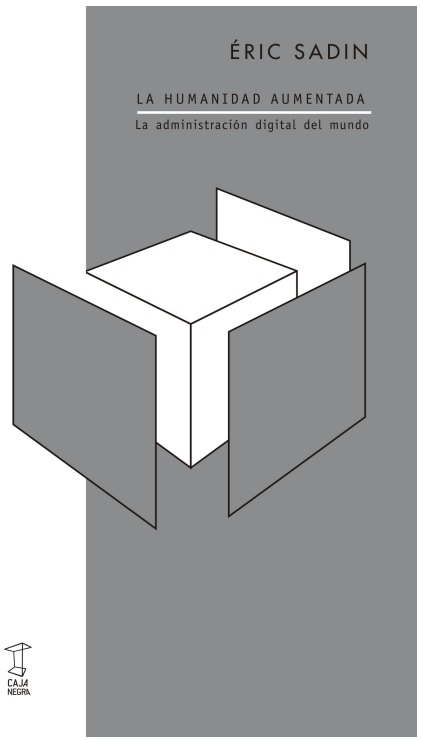
Ed. del 2015

La administración digital del mundo

Hay que inscribir la enumeración anterior en:

- una administración robotizada de la existencia humana,
- una duplicación digital del mundo,
- una administración digital del mundo y,
- en una mutación decisiva en nuestro vínculo con la técnica

en la perspectiva que plantea Eric Sadin (2018), cuyo texto fue traducido por Javier Blanco.



Las culturas generativa y predictiva

Si los datos vienen generados por:

en relación al análisis de datos, Breiman[4] distingue dos objetivos:

- Predecir cuál será el valor de la respuesta Y para valores futuros de $\boldsymbol{X}$.
- Inferir cómo la naturaleza está asociando la variable respuesta Y a las variables de entrada $\boldsymbol{X}$.

[4] Breiman, L. (2001), "StatisticalModeling: the Two Cultures", Statistical Science, 16, 199–231.

Las culturas generativa y predictiva

- En la cultura generativa:
  - se asume que hay un verdadero modelo generando los datos y,
  - el análisis de datos requiere de la inferencia.

- En la cultura predictiva[5]
  - el modelo dice poco del mecanismo subyacente que genera los datos y,
  - se hace foco en la precisión de la predicción (machine learning).

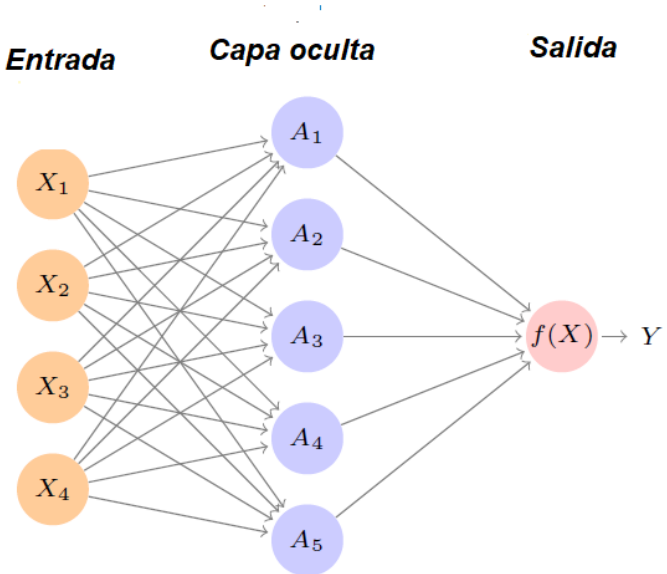
[5] Maldonado, J., Picco, M. Ruiz, M. (2022). Comparación de métodos de clasificación desde la perspectiva de la predicción y la selección del modelo. Ponencia en el Congreso de la SAE

Las culturas generativa y predictiva

- En la cultura generativa:

$$Y = \sum_{i=1}^{p} \beta_i X_i + \epsilon, \ \epsilon \sim N(\mu, \sigma^2).$$

- En la cultura predictiva:



Predicción con Random Forest

Aprendizaje supervisado. Los sensores de imágenes hiperespectrales proveen cientos de anchos de banda del espectro electromagnético de la misma área sobre la superficie terrestre.

- A cada pixel le corresponde $\boldsymbol{X} = (X_1, \ldots, X_p)$; cada entrada X_i expresa la reflectancia en una longitud de onda específica.
- Los niveles de la variable respuesta: agua, cemento y dos tipos de vegetación.
- El objetivo: entrenar un clasificador de tal modo que, a partir de un nuevo dato, x_{nuevo}, prediga (clasifique) bien.

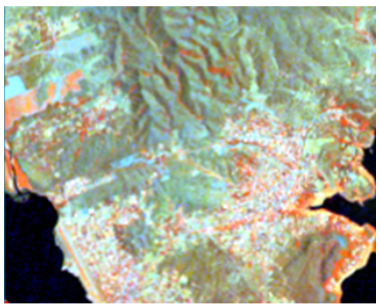
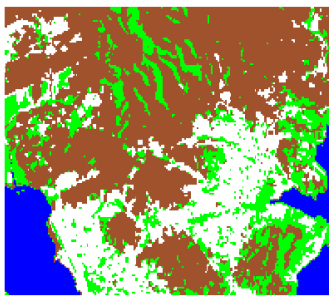


Imagen real. Hyperion EO-1.

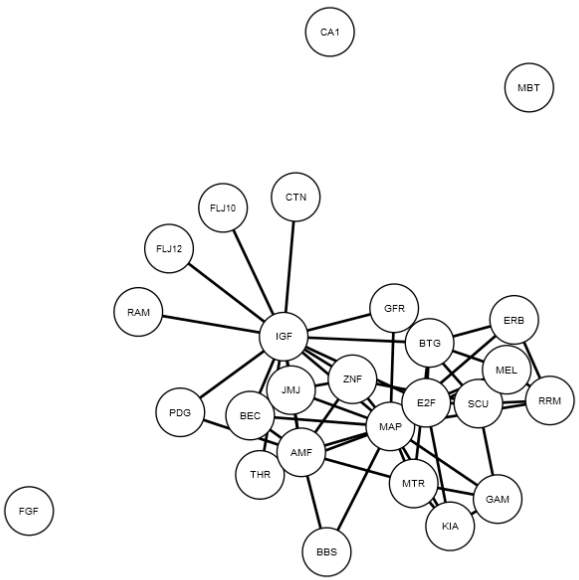


Mapa temático.
Random forest.

Inferencia con Modelos gráficos

Aprendizaje no supervisado. Ciertos tipos de cáncer quedan caracterizados por perfiles génicos:

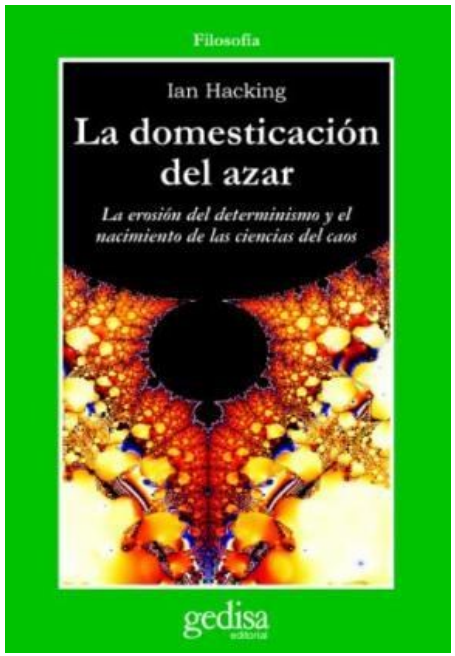
- Cada entrada, X_i, del vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ representa la expresión de la actividad del gen i, $i = 1, \ldots, p$.
- Interesa estimar la asociación de dos genes condicional a los restantes genes, para todos los pares de genes posibles.
- Si dos genes están relacionados condicionalmente a los restantes ponemos un lado el grafo.
- El objetivo: inferir el grafo asociado a $\boldsymbol{X}$.



Grafo estimado para 26 genes asociados al cáncer de mama de pacientes con respuesta patológica completa.

Alta dimensionalidad

- ¿Los "Big data" caracterizan el surgimiento de la CD?
- Los datos censales surgieron siglos atras y son big data.
- En el análisis de datos clásico n es grande en relación al número de variables p.



¡Muy buen texto de historia!

Alta dimensionalidad

- En las últimas décadas, la novedad práctica y teórica es

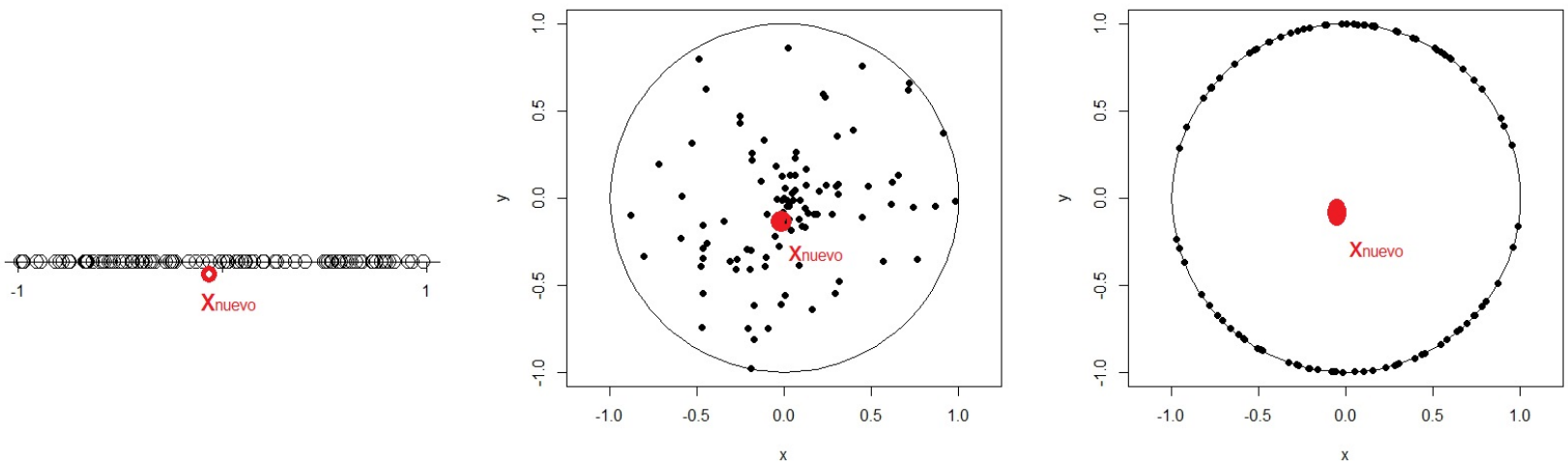
$$p >> n \text{ ó } p \approx n$$

$$\mathbb{X} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_p \\ \hline \text{caso } 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \text{caso } 2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{caso } n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}\right)$$

  y $p \to \infty$.
- En genómica, el número de genes $p = 20000$ y $n = 80$.
- En las redes sociales, el número de usuarios puede ser mucho menor que el número de variables que escoge.
- En teledetección, en imágenes hiperespectrales $p \approx n$.

Alta dimensionalidad

En alta dimensionalidad los los puntos están aislados



Entornos para $p = 1, 2$ y p "grande".

Datos administrativos y transaccionales

David Hand[6] puso de relevancia aspectos importantes de la recolección y análisis de los datos administrativos (DA).

- Los DA se recolectan por sistemas administrativos incluyendo el sector público.
- Son generados durante el curso de alguna operación y retenidos en una base de datos.
- Los DA han sido principales impulsores del desarrollo de la tecnología de minería de datos y de la CD.

[6] Hand, D. (2018). Statistical challenges of administrative and transaction data. J. R. Statist. Soc. A 181, Part 3, pp. 555–605

Datos administrativos y transaccionales

Ejemplos de DA[7]:

- Registros educativos
- Registros hospitalarios tomados de la visita de pacientes.
- Datos de los clientes de entidades financieras y bancarias.
- Información recolectada en forma automática de las prestaciones en obras sociales.

[7] Sitio oficial de la Unión Europea:
`https://ec.europa.eu/eurostat/cros/content/administrative-data-0_en`

Ejemplos de DA

Algunas características:

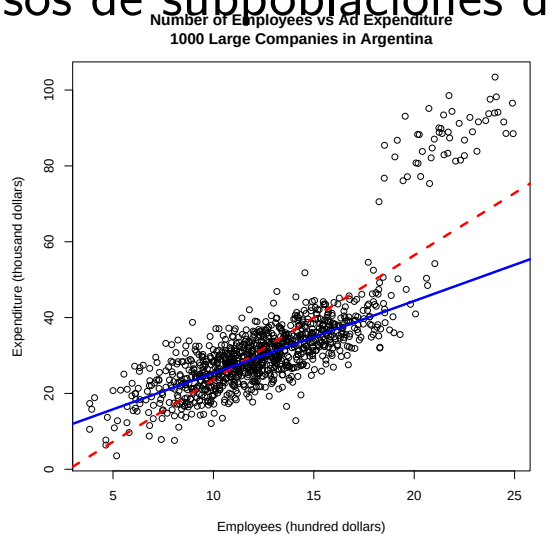
- No son muestras estadísticas, suele ser relevada la población completa.
- Se necesitan medidas descriptivas (funcionales).
- Los DA necesitan ser
  - resumidos
  - descriptos
  - analizados

  para descubrir diferentes niveles de complejidad posiblemente ocultos en diferentes niveles de complejidad y dimensionalidad.
- En general no hay problemas con la cantidad de observaciones sino con la calidad.

Análisis de datos robusto

Los DA pueden estar ingresados incorrectamente, conteniendo:[8].

- valores atípicos por celda; p. ej. 11.1 en lugar de 1.11
- filas completas atípicas; p. ej., una muestra de sangre contaminada o la inclusión de casos de subpoblaciones diferentes.



Small example to illustrate performance of robust regression algorithms.

[8] Ruiz, M., Yohai, V., Zamar, R. (2018). Discussion of: "Statistical challenges of administrative and

Algunas tareas que se realizan en Ciencia de Datos

Donoho (2015) enumera grupos de tareas:

- Recolección,preparación y exploración.
  - Recolección: desde el clásico diseño experimental a las técnicas actuales de manipulación[9]
  - Preparación, como curación de datos, entre otras tareas.
  - Exploración de datos, en el sentido clásico.
- Representación y transformación de los datos.
  - Bases de datos: implica conocer estructuras, transformaciones, algoritmos involucrados.
  - Representaciones matemáticas: como transformada de Fourier, wavelets, multiescala (en deep learning).
- Cómputos. Incluye aportar nuevas bibliotecas a lenguajes de comunidades (como R o Python), utilización de clusters.

[9] web scraping, Pubmed scraping,38 image processing, andTwitter, Facebook, and Redditmunging

¿Qué tareas se realizan en Ciencia de Datos?

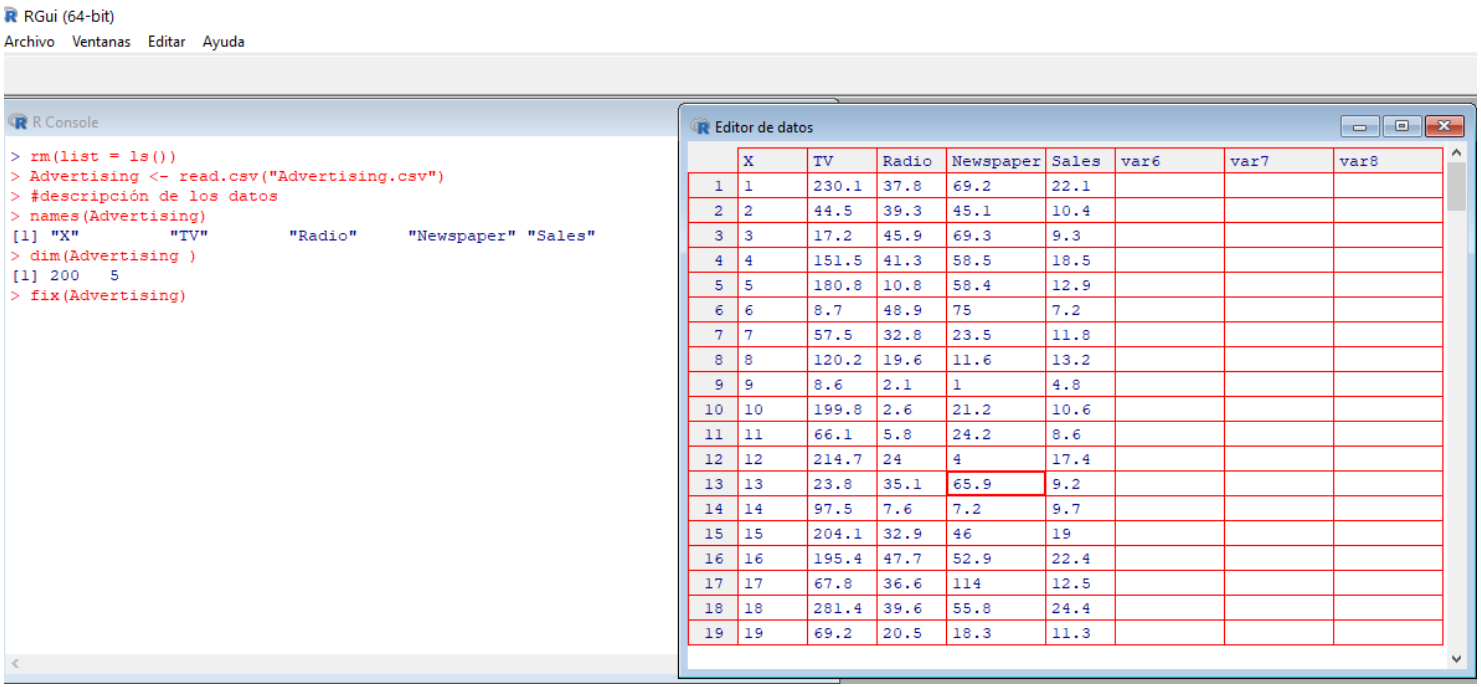
- Visualización y presentación. Va desde histogramas, boxplots, bagplots hasta dispositivos más complejos como cellhandlers, etc.
- Modelización de los datos.
  - Modelización generativa: es la estadística matemática académica tradicional.
  - Modelización predictiva: casi coincidente con el AA (machine learning).

Areas y subáreas de Estadística

- Metodología estadística general.
  - Estimación
  - Inferencia
  - Selección de modelos
  - Análisis multivariado.
  - Métodos no paramétricos.
- Bioestadística y estadística del ambiente.
- Estadística en las ciencias físicas y en la industria.
- Estadística en economía y ciencias sociales.
- Estadística computacional y CD.
  - Algorítmica
  - Grandes datos (Big data)
  - Clasificación
  - Clustering
  - Análisis de datos.
  - Minería de datos
  - Visualización
  - Análisis de imágenes
  - Aprendizaje automático
  - Datos de alta dimensión
  - Aprendizaje estadístico
  - Lenguajes y sofware
  - Pattern recognition

Problema motivador: estudio de mercado

Datos "Advertising"



	X	TV	Radio	Newspaper	Sales	var6	var7	var8
1	1	230.1	37.8	69.2	22.1			
2	2	44.5	39.3	45.1	10.4			
3	3	17.2	45.9	69.3	9.3			
4	4	151.5	41.3	58.5	18.5			
5	5	180.8	10.8	58.4	12.9			
6	6	8.7	48.9	75	7.2			
7	7	57.5	32.8	23.5	11.8			
8	8	120.2	19.6	11.6	13.2			
9	9	8.6	2.1	1	4.8			
10	10	199.8	2.6	21.2	10.6			
11	11	66.1	5.8	24.2	8.6			
12	12	214.7	24	4	17.4			
13	13	23.8	35.1	65.9	9.2			
14	14	97.5	7.6	7.2	9.7			
15	15	204.1	32.9	46	19			
16	16	195.4	47.7	52.9	22.4			
17	17	67.8	36.6	114	12.5			
18	18	281.4	39.6	55.8	24.4			
19	19	69.2	20.5	18.3	11.3			

Figura 1: $n = 200$ observaciones correspondientes, cada una, a un producto vendido en un mercado. Para cada producto conocemos el valor de venta y la inversión publicitaria en tres diferentes medios: tv, radio y diario impreso (James et al., 2021)

Estudio de mercado

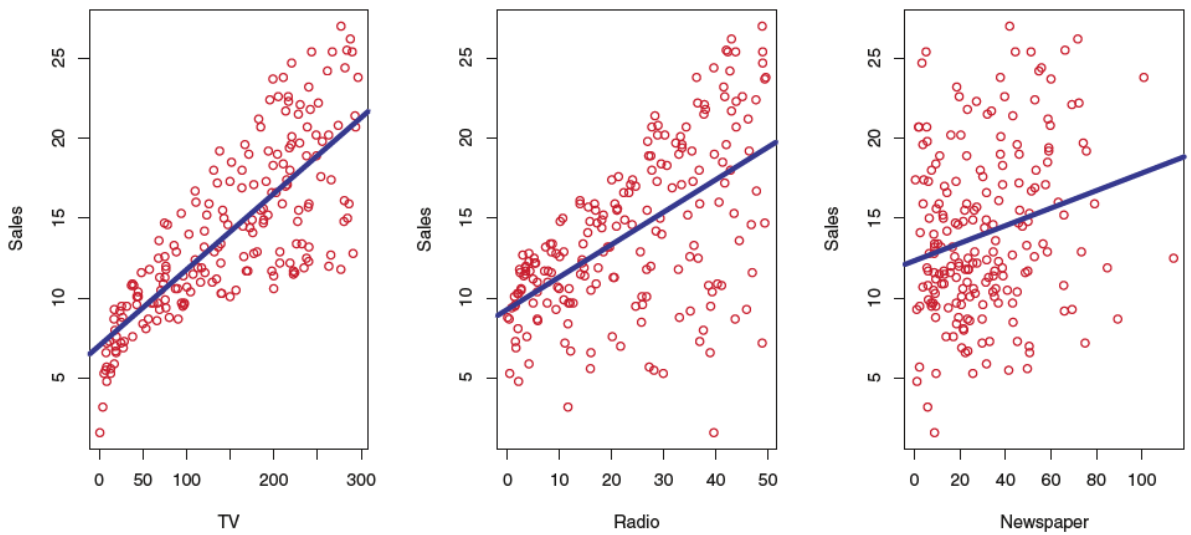


Figura 2: Datos "Advertising". Representación de ventas (sales) versus inversión según medio publicitario (TV, Radio y Newspaper)

Estudio de mercado

- No se pueden incrementar el valor de venta del producto pero sí se puede controlar el gasto en publicidad.
- Si determinamos que existe un asociación entre publicidad y ventas entonces podríamos recomendarle al vendedor cómo ajustar los montos publicitarios.

Objetivo

Desarrollar un modelo que pueda ser utilizado para predecir ventas sobre la base de tres presupuestos en medios

Estudio de mercado

Variables del modelo

- *Variables de entrada (inputs), características (features) o predictores:* "presupuesto" en TV, Radio y Newspaper.
- *Variable output, de salida, respuesta o variable dependiente*: "valor de venta" o sales.

Modelo de regresión

Más formalmente, contamos con una variable aleatoria cuantitativa Y, denominada respuesta, y un vector de predictoras[10] $\boldsymbol{x} = (X_1, \ldots, X_p)$ tal que

$$Y = f(\boldsymbol{x}) + \epsilon \tag{1}$$

donde ϵ, una variable aleatoria, representa el error, del cual asumimos que posee media cero y es independiente de $\boldsymbol{x}$.
En esta formulación f representa la *información sistemática* que $\boldsymbol{x}$ posee acerca de Y.

[10] Puede ser aleatorio o controlado

¿Por qué estimar f?

¿Por qué estimar f?

Dos razones:

- predicción
- inferencia

Predicción

Predicción

Se cuenta con un conjunto de inputs $\boldsymbol{x}$ pero el output no puede ser obtenido fácilmente.

Ejemplo

- $X_1, \ldots, X_p$ representan características de la sangre (muestra) de un paciente e Y es la respuesta, adversa, a un medicamento.
- Y no se podrá obtener por experimentación si la adversidad es severa.
  Entonces queremos estimar Y utilizando las características.

Predicción

$\widehat{Y}$

- Como $\mathrm{E}(\epsilon) = 0$, podemos predecir Y utilizando

$$\widehat{Y} = \widehat{f}(\boldsymbol{x})$$

con $\widehat{f}$ un estimador de f.

- En este contexto, no nos interesa la forma exacta de $\widehat{f}$ (caja negra) sino con qué exactitud predice Y.

Predicción

La precisión depende de dos cantidades:

- *Error reducible*: asociado al tipo de estimador $\widehat{f}$.
- *Error irreducible*: asociado al error ϵ y no depende de la precisión de $\widehat{f}$. Este error se debe a que puede contener variables no medidas.

Asumiendo $\boldsymbol{x}$ y $\widehat{f}$ fijos:

$$\mathrm{E}(Y - \widehat{Y})^2 = \underbrace{|f(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})|^2}_{\text{componente reducible}} + \underbrace{\mathrm{VAR}(\epsilon)}_{\text{componente irreducible}} \tag{2}$$

Inferencia

Objetivo

Comprender la relación existente entre las variables input y output.

Interrogantes

- ¿Qué predictores están asociados con qué respuestas?
- ¿Cuál es la relación entre la respuesta y cada predictor? Algunos predictores pueden provocar incrementos en Y y otros lo contrario, etc.
- ¿Puede la relación entre la respuesta y las predictoras resumirse adecuadamente en forma lineal o es más compleja?

Ejemplos

Predicción

Una compañía lanza una campaña de marketing directo.

- El objetivo es identificar las respuestas positivas a invitaciones por correo electrónico, basados en variables demográficas identificadas sobre cada individuo.
- No hay interés en comprender en profundidad la relación entre la respuesta a la campaña y las variables medidas sobre cada individuo.
- La compañía está interesada en obtener un modelo preciso para predecir la respuesta utilizando predictores.

Ejemplos

Inferencia

En el contexto de la base de datos "Advertising" el interés está centrado en responder a las siguientes preguntas:

- ¿Qué medio contribuye a las ventas?
- ¿Cuáles medios generan el mayor impulso de venta?
- ¿Cuál es el aumento en las ventas asociado a un incremento otorgado al presupuesto en TV?

¿Cómo estimar f?

Consideremos n realizaciones del modelo (1):

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \ i = 1, \ldots, n \tag{3}$$

donde $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ es la i-esima observación para la $j-$ésima predictora, $i = 1, \ldots, n$ y $j = 1, \ldots, p$.
La matriz de datos es

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \ldots & x_{1p} \\ y_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots & \\ y_n & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \tag{4}$$

¿Cómo estimar f?

Métodos de estimación de f

- Métodos paramétricos.
- Métodos no-paramétricos.

Métodos paramétricos

En estos métodos hay dos etapas

a) Seleccionamos un modelo, por ejemplo lineal:

$$f(\boldsymbol{X}) = \sum_{i=1}^{p} \beta_i X_i,$$

donde β_i son los coeficientes del modelo.

b) *Ajustamos* o *entrenamos* el modelo: utilizamos los datos para estimar los coeficientes.

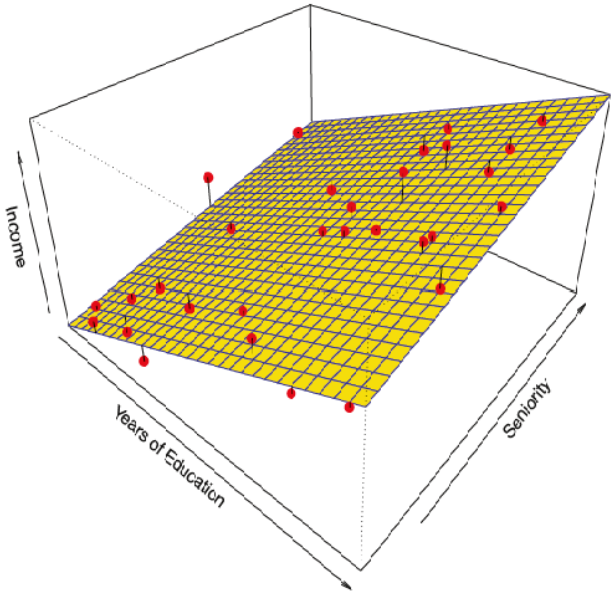
Métodos paramétricos

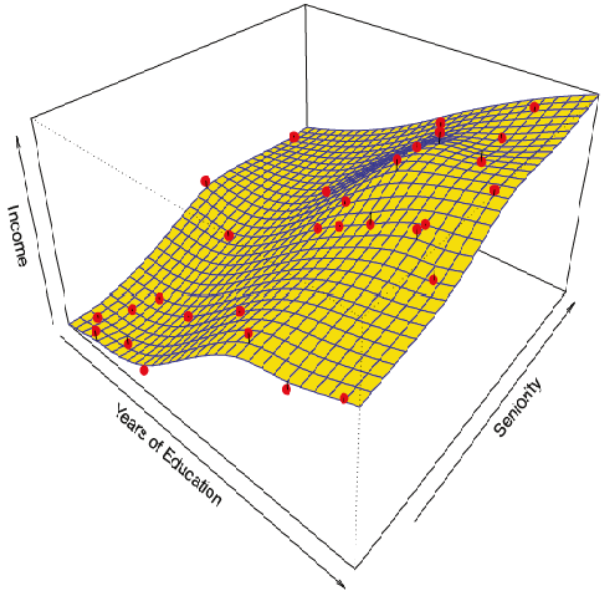
Tensiones

- Si f no es lineal, el ajuste será pobre.
- Si utilizamos modelos más flexibles (polinomios de orden superior por ejemplo) pagamos el precio de
  - estimar más parámetros;
  - posiblemente, sobreajustar los datos, *siguiendo al error demasiado cercanamente*.

Métodos paramétricos

Ajuste de los datos de ingreso versus escolaridad y antigüedad para $n = 30$ personas (¿cuál es un mejor ajuste?).





Métodos no–paramétricos

No se asume una forma particular para f.

- Las suposiciones son del tipo: f pertenece a la familia de las funciones continuas o a la de las funciones diferenciables, etc.
- Observamos que:
  - la ventaja consiste en la flexibilidad del método;
  - y su desventaja en que al no estimar un número finito de parámetros, más datos son necesarios para ajustar una función que pertenece a un espacio infinito-dimensional.

Métodos no–paramétricos: estimador basado en núcleos

Más formalmente, dado el modelo

$$Y = f(\boldsymbol{x}) + \epsilon \tag{5}$$

y si asumimos que f es derivable de orden 2 con f'' continua entonces un estimador consistente típico tiene la forma

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)}$$

con K un núcleo y h_n es una sucesión de aberturas de ventana.

¿Qué metodo utilizamos?

Los métodos son flexibles o restrictivos según las familias de funciones f sean grandes o pequeñas.

$$\mathcal{F}_0 = \{f : f \text{ es lineal}\} \subset \cdots \subset \mathcal{F}_0 = \{f : f \text{ es continua}\}$$

Por ejemplo:

- una regresión lineal es un método poco flexible o muy restrictivo;
- una regresión no paramétrica basada en núcleos es muy flexible.

¿Qué metodo utilizamos?

Compromiso entre precisión e interpretación

¿Por qué razón optaríamos por un método de escasa flexibilidad?

- Si estamos interesados en realizar una inferencia, un modelo restrictivo es siempre más simple de interpretar; por ejemplo un modelo lineal.
- Pero en cambio si sólo nos interesa predecir tal vez podamos optar por un método más complejo, menos interpretable pero con buena capacidad predictiva.

¿Qué metodo utilizamos?

Diferentes métodos estadísticos se distribuyen en el "sistema de coordenadas" de interpretabilidad y flexibilidad.

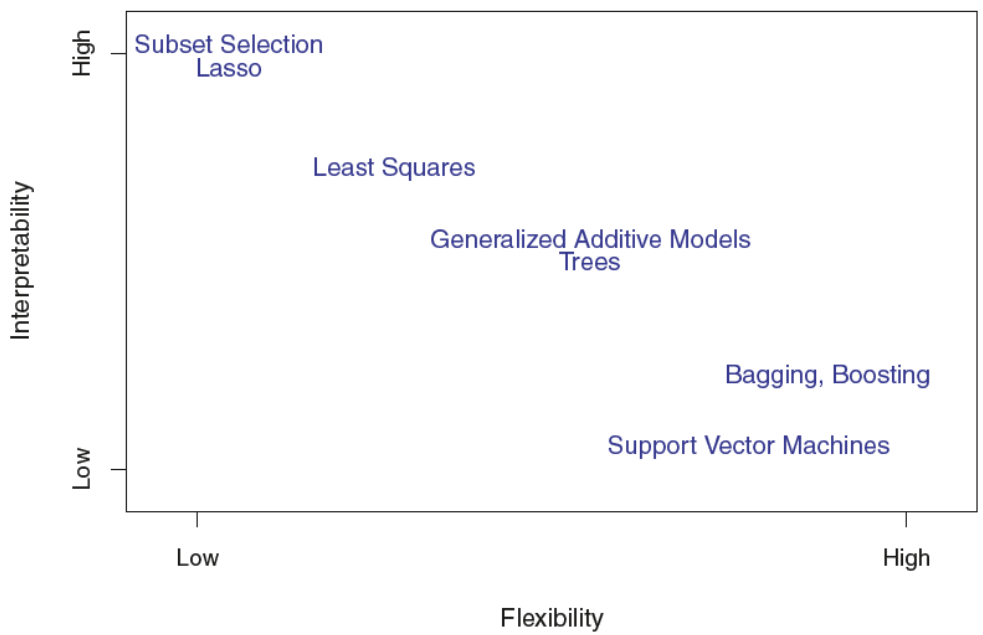


Figura 4: "Distribución" de los métodos según el compromiso entre flexibilidad e interpretación. James et al. (2021)

Aprendizaje supervisado versus no supervisado

La mayoría de los problemas estadísticos pueden agruparse en dos clasificaciones: *supervisado* y *no supervisado*.
En un problema supervisado;

- Contamos, para cada observación $\boldsymbol{x}_i$ de un predictor con una observación de una respuesta, y_i, $i = 1, \ldots, n$.
- El objetivo es ajustar un modelo que relacione la respuesta a los predictores, en vistas de predecir y/o comprender esa relación (inferencia).
- Regresión, regresión logística, modelos aditivos generalizados (GAM), boosting y support vector machines operan en el dominio del análisis supervisado.

Aprendizaje supervisado versus no supervisado

En un problema no supervisado

- contamos sólo con las observaciones del vector de predictoras $\boldsymbol{x}_i$, $i = 1, \ldots, n$, pero no tenemos respuestas,
- su denominación se debe a que carecemos de la respuesta (la variable Y) que nos pueda guiar o supervisar nuestro análisis.
- *Cluster análisis* es una técnica típica dentro de las no supervisadas.

Regresión versus Clasificación

Algunos tips

En "general", si la respuesta es cuantitativa hablamos de análisis de regresión mientras que si es cualitativa hablamos de clasificación, aunque la distinción no puede ser establecida de forma tan tajante.

Regresión versus Clasificación

Algunos tips

- La regresión por mínimos cuadrados supone una variable respuesta cuantitativa.
- La regresión logística es utilizada para una variable respuesta cualitativa con dos respuestas posibles (binaria) y, al mismo tiempo, resuelve un problema de clasificación.
- Vecinos más próximos y boosting pueden ser utilizados tanto para variables respuesta cualitativas como cuantitativas.

Precisión del modelo en el contexto de regresión

Queremos evaluar el desempeño de un método de aprendizaje estadístico sobre un conjunto de datos dados, por ende necesitamos una medida de cuán bien sus predicciones ajustan a los datos observados.
Una posible medida es el error cuadrático medio (estimado)

$$\widehat{\mathrm{ECM}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2 \tag{6}$$

donde $\widehat{f}$ es el estimador de f.

Precisión del modelo en el contexto de regresión

Es importante distinguir dos tipos de problemas:

a) Si asumimos distribución normal, independencia y que el modelo es lineal f, el problema de optimalidad de la selección de f está resuelto y hay unicidad de esa solución.

b) El escenario al que se refiere James et al. (2021) es cuando no hay cumplimento claro de supuestos y entonces aparecen varios métodos competidores de aprendizaje estadístico: regresión lineal simple, splines, regresión no paramétrica para el mismo conjunto de datos.

En ambos contextos seleccionamos aquel $\widehat{f}$ que minimice $\widehat{\text{ECM}}$[11] pero las situaciones son bien diferentes.

[11] Por simplicidad escribiremos ECM

Precisión del modelo en el contexto de regresión

Observemos que:

- El ECM está calculado sobre los datos dados, a los que llamaremos datos de entrenamiento.
- Y estamos interesados en la precisión de las predicciones que obtenemos a partir de los datos test todavía no vistos.

Datos de entrenamiento y datos test

Más generalmente, supongamos que contamos con una colección

$$C = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

de n datos de entrenamiento.
En base a C se construyó $\widehat{f}_C$ y el $\mathrm{ECM}(\widehat{f}_C)$ mide la calidad del ajuste.
Un buen método de aprendizaje estadístico es aquel que, además de minimizar el $\mathrm{ECM}(\widehat{f}_C)$, frente a un nuevo dato test $(x_0, y_0) \notin C$ satisfaga

$$y_0 \approx \widehat{f}_C(x_0).$$

Datos de entrenamiento y datos test

Y si contamos con una colección grande

$$C_N = \{(\boldsymbol{x}_1^0, y_1^0), \ldots, (\boldsymbol{x}_m^0, y_m^0)\}$$

de "nuevos" datos interesa el procedimiento $\widehat{f}_C$ que minimice tanto (6) como

$$\frac{1}{m}\sum_{i=1}^{m}(y_i^0 - \widehat{f}_C(x_i^0))^2. \tag{7}$$

A esta expresión le denominamos ECM–test o error test.

Datos de entrenamiento y datos test: ejemplo

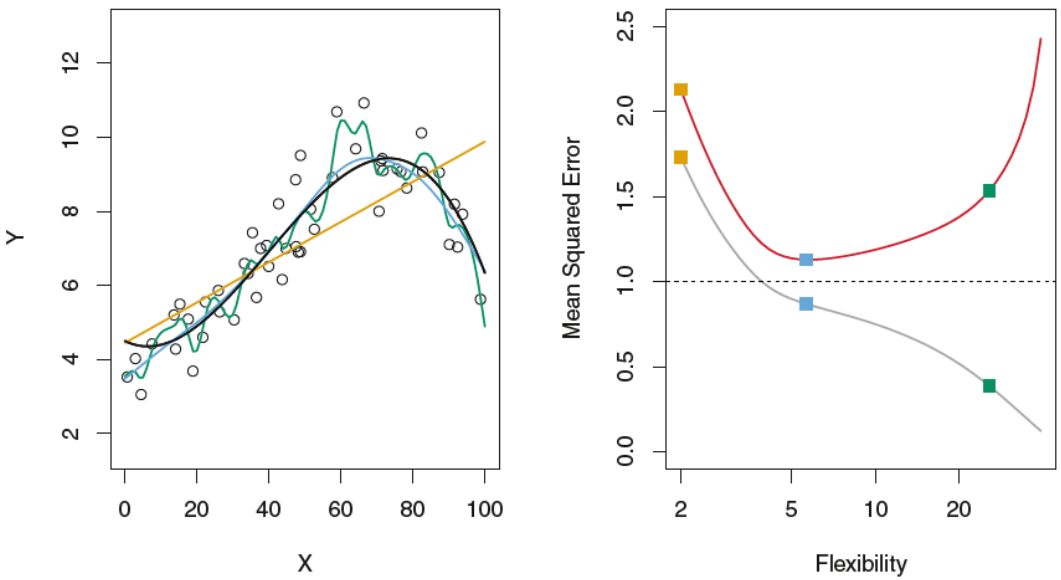


Figura 5: Desempeño de tres métodos de estimación de f.

Se generan observaciones a partir f (conocida) y se comparan tres métodos de estimación de f. En el panel izquierdo se representan: estimación por regresión lineal (línea naranja) y por splines (líneas azul y verde). La línea negra sólida corresponde a la curva de f. En el panel derecho se comparan los valores de ECM para los datos de entrenamiento (curva gris), para los datos test (curva roja) y el mínimo ECM–test sobre todos los métodos (línea quebrada). James et al. (2021)

Datos de entrenamiento y datos test: ejemplo

Observaciones. En relación al panel izquierdo de la Figura 5.

- Las líneas naranja, azul y verde representan tres estimadores para f, utilizando métodos con niveles de flexibilidad creciente.
- La línea naranja indica un ajuste por regresión lineal, relativamente inflexible. La azul y verde fueron producidos por splines con diferentes niveles de suavizado.
- A medida que aumenta el suavizado las curvas de los estimadores se aproximan mejor a los datos.
- La curva verde es la más flexible y se parea con los datos muy bien pero ajusta pobremente a f por su alta variabilidad.

Datos de entrenamiento y datos test: ejemplo

Observaciones. Panel derecho de la Figura 5.

- La línea gris muestra el ECM basado en los datos de entrenamiento como una función de la flexibilidad. Los cuadrados naranja, azul y verde indican los valores de ECM asociados con las correspondientes curvas del panel izquierdo.
- Utilizaremos los términos **"grados de libertad"** para indicar, provisoriamente, **"flexibilidad de una curva"**. Cuanto mayor ondulaciones tenga una curva más grados de libertad posee.

Datos de entrenamiento y datos test: ejemplo

Observaciones. Panel derecho de la Figura 5.

- El ECM-entrenamiento decrece monótonamente cuando la flexibilidad aumenta.
- Como f no es lineal, el "ajuste naranja" es pobre mientras que la curva verde muestra el más bajo de los ECM-entrenamiento por su alto grado de libertad.

Datos de entrenamiento y datos test: ejemplo

Observaciones. En relación al panel derecho de la Figura 5.

- La línea roja muestra el ECM basado en los datos test.
- Similar a ECM-entrenamiento, ECM-test decrece inicialmente cuando aumenta la flexibilidad pero a partir de cierto punto comienza a aumentar.
- Así, las curvas naranja y verdes tienen peor desempeño que la azul (mayor ECM-test).
- La línea horizontal quebrada indica $\mathrm{VAR}(\epsilon)$, el componente irreducible de (2) y por ende el mínimo ECM-test alcanzable entre todos los métodos posibles.

Más allá del ejemplo

Dimensiones generales del entrenamiento estadístico

- El decrecimiento ECM-entrenamiento a medida que la flexibilidad aumenta y la forma de U de ECM-test versus flexibilidad es un fenómeno general de los **métodos de aprendizaje estadístico** (mae).
- Cuando la flexibilidad aumenta ECM-entrenamiento decrerá pero puede que ECM-test no.

Más allá del ejemplo

Dimensiones generales del entrenamiento estadístico

Cuando un método produce un ECM-entrenamiento muy pequeño y un ECM-test grande estamos frente a un sobreajuste de los datos que:

- se debe a que nuestro mae trabaja demasiado exhaustivamente para encontrar patrones en los datos de entrenamiento y puede que halle algunos patrones que se deben sólo al fenómeno aleatorio y no a las propiedades verdaderas de f;
- provoca un ECM-test grande, como efecto de que los patrones que el mae halla en los datos de entrenamiento simplemente no existen en los datos test.

Más allá del ejemplo

Dimensiones generales del entrenamiento estadístico

- En general

$$\text{ECM-entrenamiento} < \text{ECM-test}.$$

.

- Cuando hablamos de sobreajuste de un mae es porque hay otro mae menos flexible que ha producido ya un ECM-test menor.

Más allá del ejemplo

- En muchos contextos podemos contar con observaciones test, pero en general no
- ¿Y si sólo contamos con los datos de entrenamiento? Hay que construir una estrategia de expresión máxima de los datos de tal modo que el método aprenda (convalidación o validación cruzada, por ejemplo).

Descomposición sesgo-varianza

Se puede demostrar que, para el modelo (1), en condiciones muy generales

$$\begin{aligned} \text{ECM}\left(\hat{f}\left(x_0\right)\right) &=: \mathsf{E}\left(y_0 - \hat{f}\left(x_0\right)\right)^2 \\ &= \text{VAR}\left(\hat{f}\left(x_0\right)\right) + \left[\text{SES}\left(\hat{f}\left(x_0\right)\right)\right]^2 + \text{VAR}(\epsilon) \end{aligned} \quad (8)$$

donde

$$\text{SES}\left(\hat{f}\left(x_0\right)\right) = \mathsf{E}(\hat{f}\left(x_0\right)) - f(x_0)$$

es el sesgo del estimador.

Descomposición sesgo-varianza

Observaciones importantes:

- El ECM global se obtiene promediando (computo una integral o una serie) $\mathrm{ECM}\left(\hat{f}\left(x_0\right)\right)$ sobre todos los puntos x_0.
- $\forall x_0$:

$$\mathrm{ECM}\left(\hat{f}\left(x_0\right)\right) \geq \mathrm{VAR}(\epsilon)$$

- La igualdad (8) nos indica que un buen mae debe tener bajos sesgo y varianza.
- Un buen método debe minimizar el El ECM global.

Compromiso sesgo-varianza

Componente varianza

- Elegido el mae, diferentes estras de datos de entrenamiento producen distintos $\widehat{f}$. Un buen mae no debería variar demasiado entre las diferentes muestras. Esta es la noción de varianza del método.
- En general los mae muy flexibles tienen alta varianza.

Recordemos el panel izquierdo de la Figura 5, ahora Figura 6.

Compromiso sesgo-varianza

Componente varianza

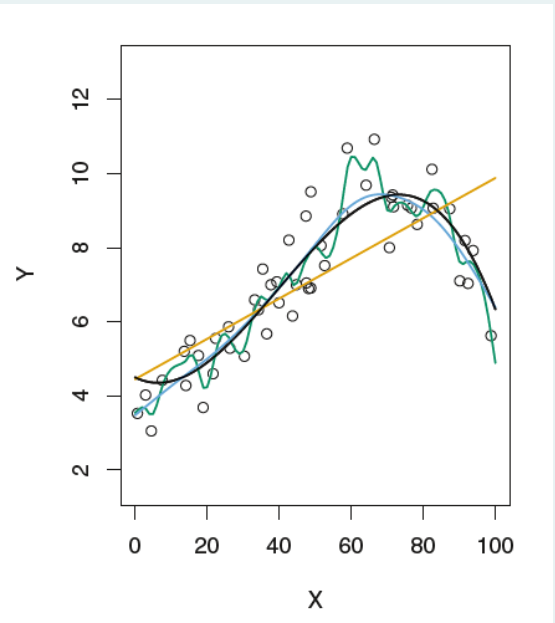


Figura 6: Desempeño de tres métodos de estimación de f. La curva verde (naranja) representa un estimador muy variable ya que sigue a los datos (poco flexible y con baja varianza).

Compromiso sesgo-varianza

Compromiso sesgo–varianza

Figura 7:

- Cuando f no es lineal, en un ajuste lineal (baja flexibilidad) el aporte al MSE lo realiza el sesgo y el error es grande. En cambio si la flexibilidad es alta el MSE es bajo y el sesgo es casi nulo.

Figura 8:

- Cuando f es casi lineal un ajuste lineal (baja flexibilidad) minimiza MSE. Un ajuste no lineal (muy flexible) provoca un sesgo casi nulo pero un alto MSE el cual se explica por la varianza.

Compromiso sesgo-varianza

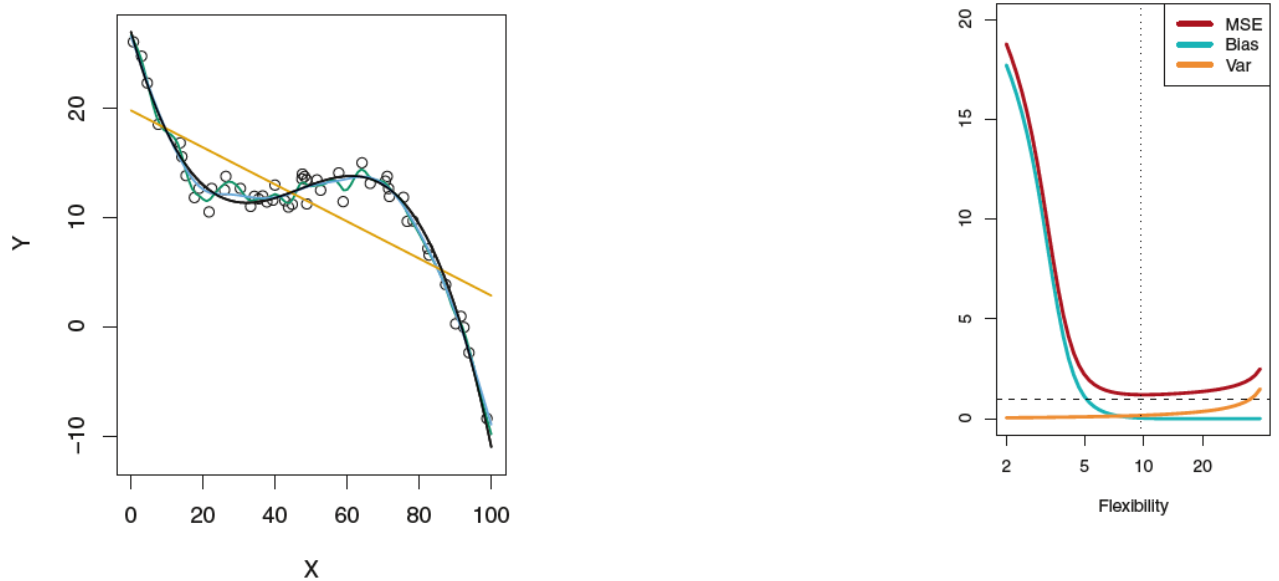
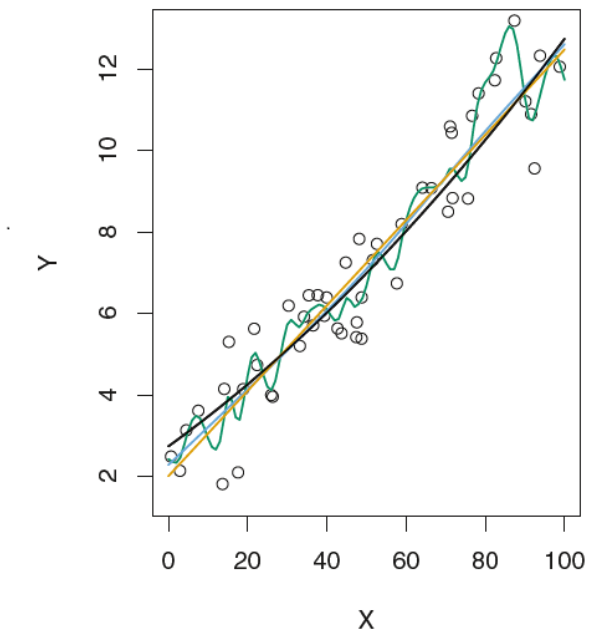


Figura 7: Contribución del sesgo y la varianza cuando f no es lineal

Compromiso sesgo-varianza



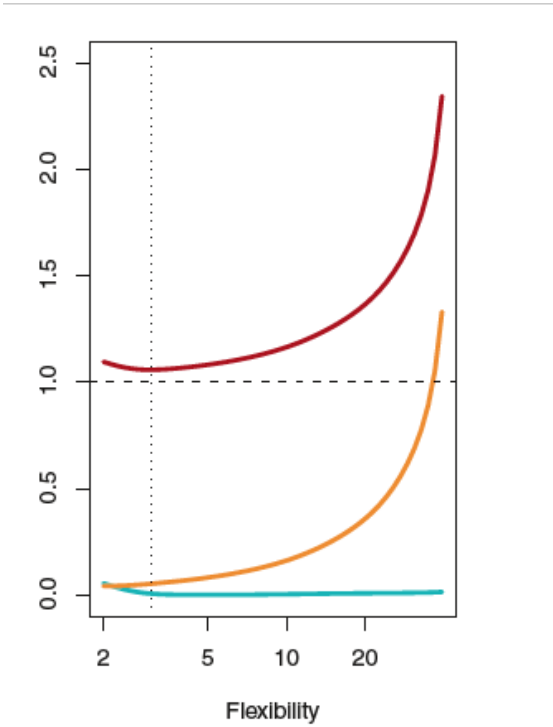


Figura 8: Contribución del sesgo y la varianza cuando f es casi lineal

Clasificación

Como antes consideremos n observaciones de $(\boldsymbol{x}, Y)$, $(\boldsymbol{x}_1, y_1)$, ..., $(\boldsymbol{x}_n, y_n)$, donde la variable respuesta Y es cualitativa.
Ejemplos típicos:

- Una persona arriba a un centro de emergencia con un conjunto de síntomas (predictoras) que podrían ser atribuidos a tres condiciones médicas (respuesta).
- Una línea de servicio de un banco determina cuando una transacción es fraudulenta de acuerdo a dirección IP del usuario, historia pasada de las transacciones, etc.
- De acuerdo al largo y ancho del sépalo y del pétalo se clasifica la especie de una flor Iris.

Clasificación

Nuestro estimador $\widehat{f}$ es un clasificador (regla).

Tasa de error de entrenamiento

- La precisión de $\widehat{f}$ es la tasa de error de entrenamiento: la proporción de errores que se producen cuando aplicamos $\widehat{f}$ a los datos de entrenamiento; i.e.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i)$$

donde $\widehat{y}_i$ es la etiqueta de clase predicha para la i-ésima observación utilizando $\widehat{f}$.

Clasificación

Nuestro estimador $\widehat{f}$ es un clasificador (regla).

Tasa de error test

- La *tasa de error test* se define de manera análoga para un conjunto de observaciones test. Más específicamente, dado $(\boldsymbol{x_0}, y_0)$, esa tasa es

$$\text{promedio}(I(y_0 \neq \widehat{y}_0)).$$

- *Un buen clasificador es aquel que hace mínima esta tasa de error test.*

Clasificador de Bayes

La *tasa de error test* es minimizada, en promedio, por el clasificador de Bayes que *asigna a cada observación la clase más probable dado los valores del predictor*.
Formalmente, dado $\boldsymbol{x} = \boldsymbol{x}_0$ le asignamos $Y = j_0$ sii

$$P(Y = j_0 | \boldsymbol{X} = \boldsymbol{x}_0) = \max_j P(Y = j | \boldsymbol{X} = \boldsymbol{x}_0).$$

Así, este clasificador es el mejor u óptimo.

Clasificador de Bayes

Clasificador de Bayes con dos clases 1 y 2

Dado $\boldsymbol{x} = \boldsymbol{x}_0$, el clasificador de Bayes corresponde a predecir la clase 1 si $P(Y = 1|\boldsymbol{x} = \boldsymbol{x}_0) > \frac{1}{2}$ y, la clase 2, si es de otro modo.

Frontera de decisión de Bayes

se define como el conjunto de puntos

$$\left\{ \boldsymbol{x} : P(Y = 1|\boldsymbol{x} = \boldsymbol{x}) = P(Y = 0|\boldsymbol{x} = \boldsymbol{x}) = \frac{1}{2} \right\}.$$

k-vecinos más próximos

- El clasificador de Bayes, que es el optimal, supone conocidas las probabilidades de clasificación

$$P(Y = j_0|\mathbf{X} = \mathbf{x}_0), \forall j, \forall \mathbf{x}_0.$$

- Pero para un conjunto de datos reales (que no sean simulados de una distribución conocida) estas probabilidades no se conocen y se deben estimar.

k-vecinos más próximos

El estimador de *"$K-$vecinos más próximos"* (KNN) procede por estimación de dichas probabilidades.
Construyamos el estimador por pasos:

Paso 1 Dado un entero positivo K y una observación test $\mathbf{x}_0$, el KNN identifica los K puntos en la muestra de entrenamiento más cercanos (utilizando alguna métrica) a $\mathbf{x}_0$. Denotamos con $\mathcal{N}_{\mathbf{x}_0}$ al conjunto de los vecinos.

k-vecinos más próximos

Paso 2 Estimemos la probabilidad condicional $\pi_{j|\mathbf{x}_0} = P(Y = j_0|\mathbf{X} = \mathbf{x}_0)$ como

$$\widehat{\pi}_{j|\mathbf{x}_0} = \frac{1}{K} \sum_{i \in \mathcal{N}_{\mathbf{x}_0}} I(y_i = j)$$

Paso 3 El clasificador KNN aplica la regla de Bayes: a la observación $\mathbf{x}_0$ le asigna la clase j_0 que maximiza $\widehat{\pi}_{j|\mathbf{x}_0}$ en la colección de todas las clases.

k-vecinos más próximos

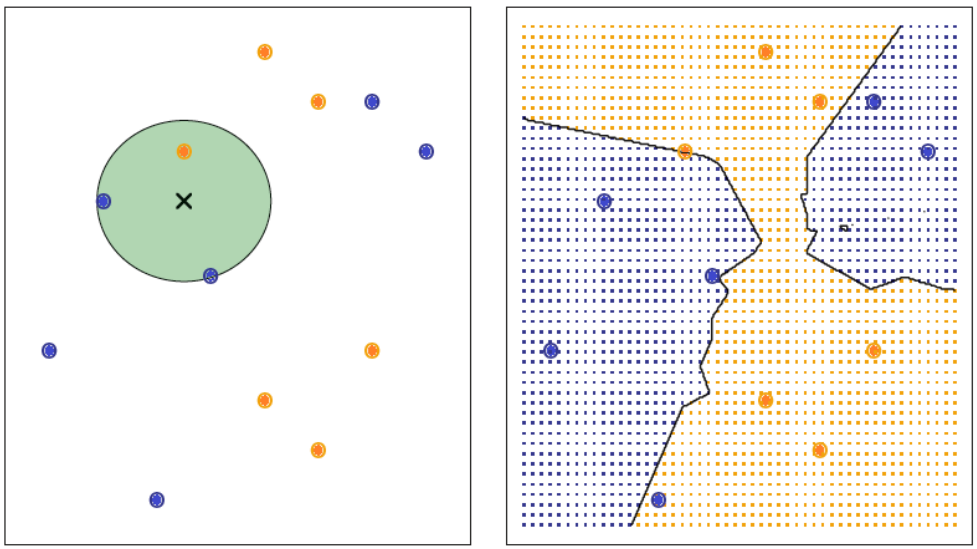


Figura 9: KNN con $K = 3$ vecinos más próximos[12]. En el panel izquierdo se exhibe la nueva observación x clasificada como azul. En el panel derecho se representa la frontera de decisión y las clases.

[12] Tomada de James et al. (2021)

Introducción: datos de alta dimensión

Son consecuencia del registro simultáneo de miles (o millones) de características sobre cada objeto o individuo

- Las biotecnologías asociadas a la genética como los microarreglos.
- Imágenes y videos: imágenes médicas, astrofísicas, de vigilancia, etc.
- Datos de preferencias de consumo: programas de fidelización, redes sociales, etc.
- Datos provistos por los negocios: compañías de logística y transporte, de seguros, de la industria financiera, etc.
- Datos colaborativos:los datos surgen de los registros de miles de usuarios a través de celulares. Ejemplo: "eBirth".

¿Bendición o maldición?

¿No es acaso, para la estadística, una bendición poder registrar miles de variables para un individuo?

La respuesta es... ¡depende!

En estadística de alta dimensión separar en general la señal del ruido es casi imposible

La maldición de la dimensionalidad

El impacto de una dimensión alta puede ser enorme y paradojal. En $\mathbb{R}^p$ con p grande

- Los puntos son aislados
- Las acumulaciones de pequeñas fluctuaciones en muchas direcciones diferentes pueden producir una gran fluctuación global
- Un evento que es acumulación de eventos raros puede no ser raro
- Los cómputos numéricos y las optimizaciones pueden tener alto costo

Los espacios son vastos

Modelo de Regresión

$$Y_i = f(\boldsymbol{X}^{(i)}) + \epsilon_i \tag{9}$$

con

- $\boldsymbol{X}^{(i)}$, $i = 1, \ldots, n$ observaciones i.i.d de un vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ de dimensión p
- $\epsilon_1, \ldots, \epsilon_n$ son i.i.d y están centradas.

Estimamos $f(\boldsymbol{x})$ utilizando *información local* , por ejemplo con k-vecinos más próximos.

Si p es grande ¿cuántas observaciones hay cercanas a $\boldsymbol{X} = \boldsymbol{x}$?

Los espacios son vastos

Una pregunta previa a la anterior:

$$\text{¿Cómo se comportan las distancias } \left\| \boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)} \right\|^2 ?$$

Para simplificar $X_1, \ldots, X_p$ son i.i.d con $X_1 \sim U([0,1])$

$$\Downarrow$$

$$\boldsymbol{X} = (X_1, \ldots, X_p)' \sim U([0,1]^p)$$

Las Figuras 10 y 11 muestran los cuatro histogramas de

$$\left\| \boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)} \right\|, 1 \leq i < j \leq n = 100$$

en cuatro escenarios (dimensiones):

$$p = 2, 10, 100, 1000.$$

Los espacios son vastos

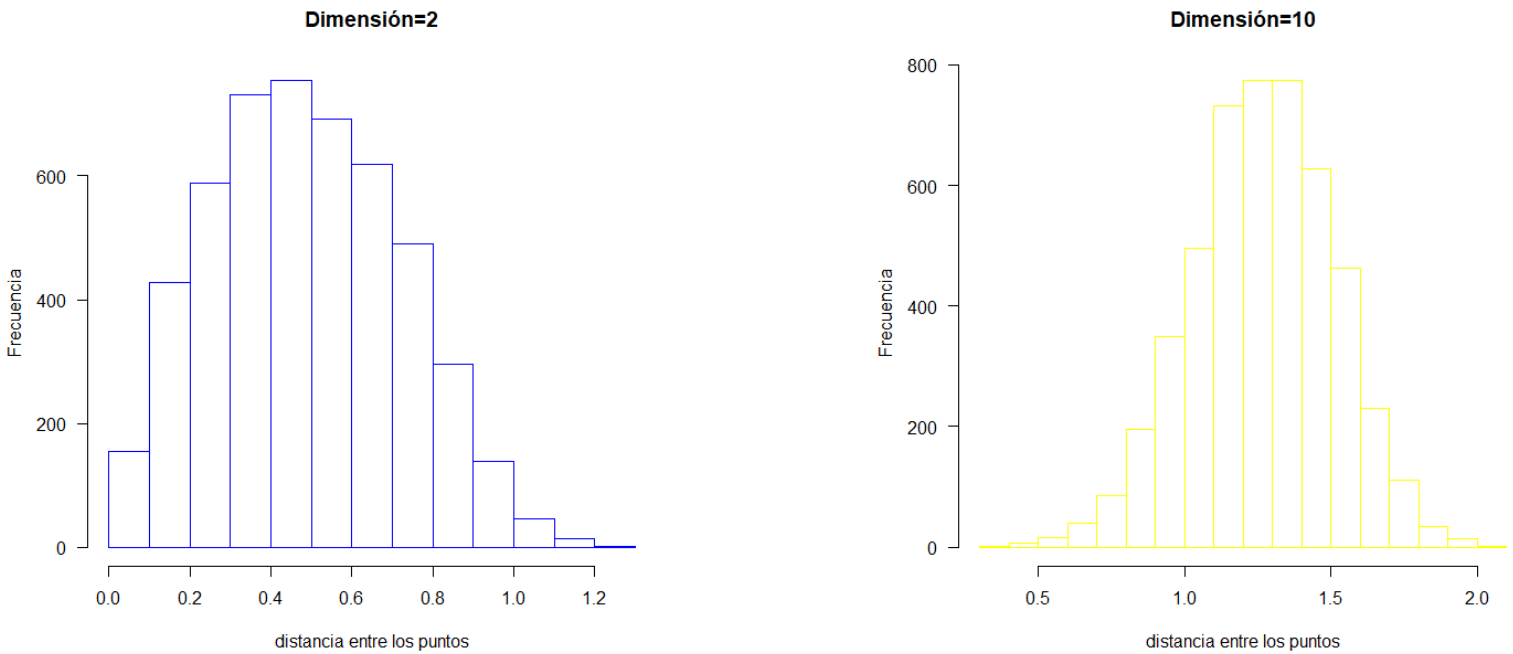


Figura 10: Distribuciones empíricas de las distancias, $p = 2, 10$.

Los espacios son vastos

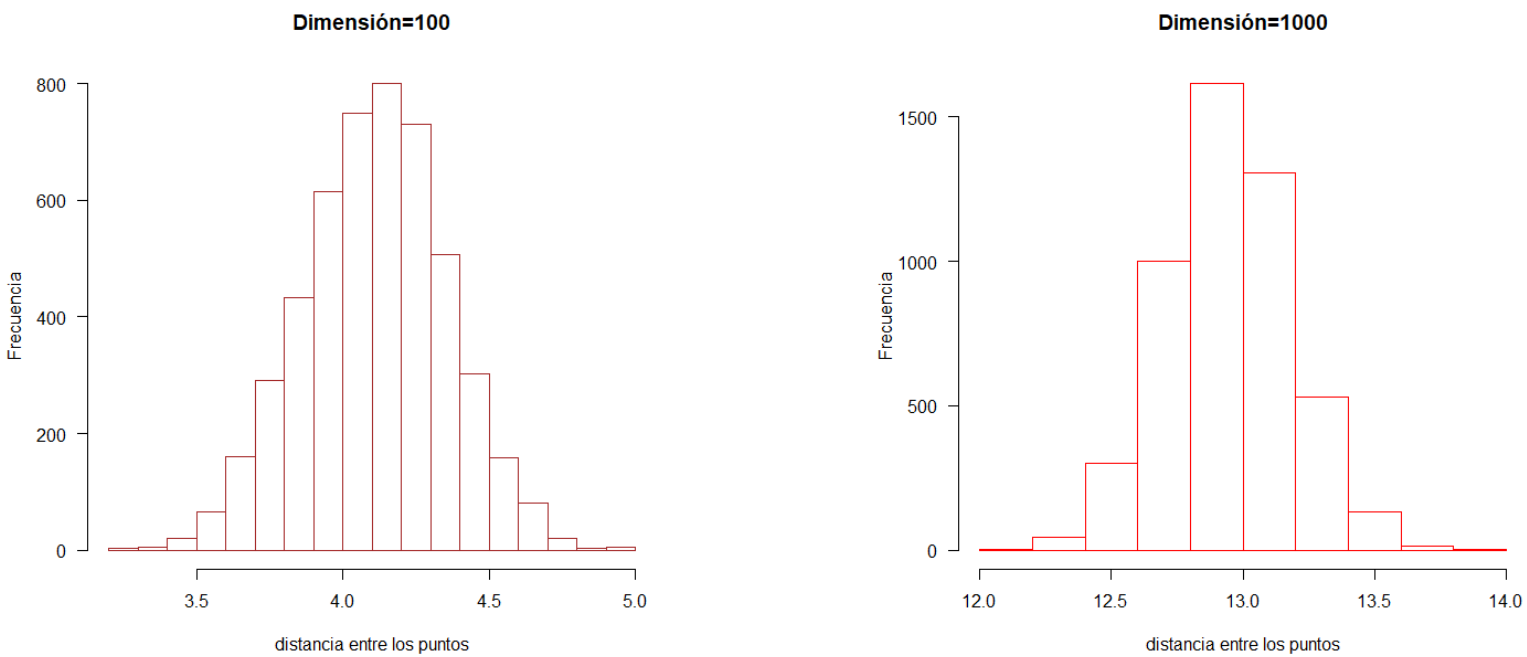


Figura 11: Distribuciones empíricas de las distancias, $p = 100, 1000$.

Los espacios son vastos

De los resultados empíricos

- La distancia entre los puntos se incrementa con p.
- Si p es grande todos los puntos están a una distancia similar.
- Notar que la DCM y el DE [13] satisfacen

$$\text{DMC} = \mathrm{E}\left\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\right\|^2 = p/6$$

$$\text{SD} = \left(\sum_{k=1}^{p} \mathrm{VAR}\left[\left(\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\right)^2\right]\right)^{1/2} \approx 0.2\sqrt{p}$$

$$\text{SD/DMC} \sim p^{-1/2}.$$

[13] Distancia Media Cuadrática y Desvío Estándar, respectivamente

Los espacios son vastos

Dado $\boldsymbol{x} \in [0,1]^p$,

$$¿n : \exists \boldsymbol{X}^{(i)} : \left\| \boldsymbol{X}^{(i)} - \boldsymbol{x} \right\| \leq 1 \ ?$$

El volumen de la bola p-dimensional de radio r es

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} r^p \quad \sim \quad \left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2} \qquad (10)$$
$$p \quad \to \quad \infty$$

Si

$$[0,1]^p \subseteq \bigcup_{i=1}^{n} B(\boldsymbol{X}^{(i)}, 1)$$

entonces

Los espacios son vastos

Volviendo a la pregunta, dado $\boldsymbol{x} \in [0,1]^p$,

$$¿n : \exists \boldsymbol{X}^{(i)} : \left\| \boldsymbol{X}^{(i)} - \boldsymbol{x} \right\| \leq 1 \ ?$$

- Si $p = 20$ entonces $n = 39$
- Si $p = 50$ entonces $n = 5.7 \times 10^{12}$
- Si $p = 200$ entonces

  n es mayor que el número total de partículas observables en el universo

Los espacios son vastos

De (10) se deduce que:

- $V_p(r) \downarrow 0$ con una tasa más rápida que una exponencial si $p \uparrow \infty$
- por ejemplo, para $p = 20$ el volumen de la bola unidad es (casi) cero!
- si $C_p(r) = \{\boldsymbol{x} \in [0,1]^p : 0.99r \leq \boldsymbol{x} < r\}$ entonces

$$\frac{\text{volumen}\left(C_p(r)\right)}{\text{volumen}\left(B_p(r)\right)} \sim 1$$

  exponencialmente cuando $p \to \infty$.
- si p es grande, "la mayoría" de los puntos de la bola $B_p(r)$ están más cerca de su frontera que de su centro.

Regresión Lineal en alta dimensión

Consideremos el siguiente modelo de regresión

$$Y_i = \left\langle \boldsymbol{X}^{(i)}, \beta^* \right\rangle + \epsilon_i, i = 1, \ldots, n,$$

con $\boldsymbol{X}$ la matriz de datos, de la que asumimos que sus columnas son ortonormales. Entonces

$$\mathrm{E}\left[\left\|\widehat{\beta} - \beta^*\right\|^2\right] = p\sigma^2$$

con lo cual si la dimensión p es grande el error es grande.

Regresión Lineal en alta dimensión

Ejemplo: Consideremos el modelo

$$Y_i = f_{\beta^*}(i/n) + \epsilon_i,\ i = 1, \ldots, n$$

siendo

$$f_{\beta^*}(t) = \sum_{j=1}^{p} \beta_j^* \cos(\pi j t)$$

y donde $\epsilon_1, \ldots, \epsilon_n$ es una muestra aleatoria de una normal estándar.
En la Figura 12 se representan 100 observaciones del modelo y las curvas correspondientes a la función de regresión y al estimador de mínimos cuadros en cuatro escenarios: $p = 10, 20, 50, 100$. Los β_j^* son constantes muestreadas de una $\mathrm{N}(0, j^{-4})$.

Regresión Lineal en alta dimensión

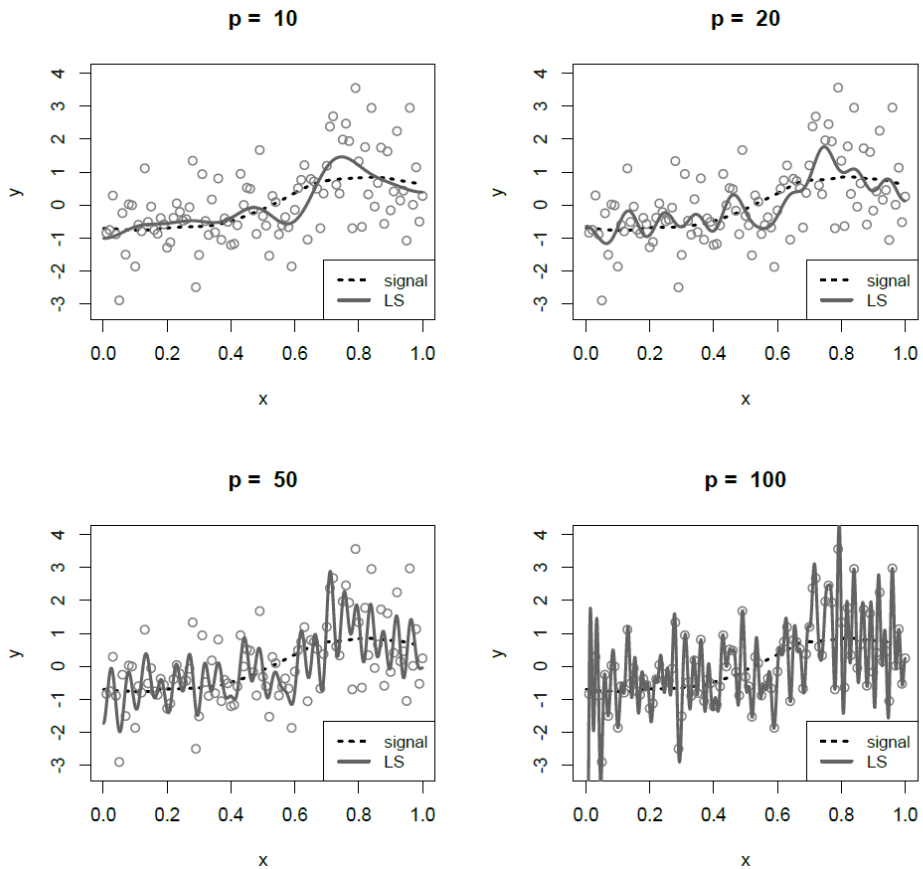


Figura 12: Estimador de mínimos cuadrados $f_{\hat{\beta}}$, con $n = 100$ y $p = 10, 20, 50, 100$. Los puntos son

Complejidad computacional

En un contexto de alta dimensión surgen

- una gigantesca intensidad del cómputo numérico
- e incluso la posibilidad de exceder las fuentes o recursos (memoria) computacionales

Cambio de paradigma

En estadística clásica hay una rica teoría para analizar los datos cuando n es grande y p pequeño.

Pero en muchos campos científicos esto no ocurre y, en cambio,

- hay un número grande de parámetros p.
- puede haber tamaños de muestras del mismo tamaño que p o incluso de tamaño menor.

En consecuencia, en este nuevo contexto,

la teoría asintótica "p fijo y $n \to \infty$" no tiene sentido.