NÚMEROS ALEATORIOS

EXPERIMENTO ALEATORIO

Ensayo que produce un resultado diferente cuando se ejecuta repetidas veces bajo las mismas condiciones iniciales.

Si bien no es posible predecir el resultad de cada ensayo, si se puede predecir el conjunto de posibles resultados y la medida de probabilidad, es decir, el espacio de probabilidad. Ej: Llegada de fotones a un CCD, dado.

NÚMEROS ALEATORIOS & PSEUDOALEATORIOS

> Sea E un experimento aleatorio y S su espacio muestral (conjunto de todos los posibles resultados de E), toda función que asigne a cada uno de los elementos de S un número real $x(s)$ se llama VARIABLE ALEATORIA (V.A)

$$X : S \to \mathbb{R}$$

Una V.A puede ser → discreta (conjunto contable).
→ continua (intervalo de números).

Los números aleatorios son observaciones independientes de V.A. distribuidas uniformemente en el intervalo (0,1).

¡PERO! en la práctica se obtienen números pseudoaleatorios: secuencia de números generados por un algoritmo que tienen la apariencia de ser V.A uniformes e independientes en (0,1). El algoritmo genera una cadena de nº que se puede predecir exactamente si se conoce dicho algoritmo.

* GENERADORES DE NÚMEROS ALEATORIOS

Los generadores deben cumplir 3 propiedades:

- Repetibilidad → si se repiten los parámetros se repite la secuencia.
- Portabilidad → bajo las mismas condiciones de definición, la sucesión es la misma (independientemente del lenguaje).
- Velocidad computacional → + rápido ≡ + resultados, conviene usar lenguajes de bajo nivel.

Además, los generadores deben ser 'intuitivamente' aleatorios.

① Generador de Von Neumann

- La semilla x_0 es un nº entero de 5 dígitos aleatorios. (03001)
- Se calcula x_0^2 y se lo escribe en 10 dígitos. (0009006001)
- El x_1 está dado por los 5 dígitos centrales. (09006)
- Se repite.

Las desventajas son que: No se conocen propiedades teóricas del generador (como la semilla conveniente) y hay malas semillas.

Scanned with CamScanner

② Generador de congruencia lineal

Genera una sucesión de enteros $y_1, \dots, y_n$ entre 0 y m-1 usando la regla recursiva:

$$y_i = (a \cdot y_{i-1} + c) \bmod [m] \qquad 0 \leq y_i \leq m-1$$

resto de la división por m.

donde: y_0 = semilla. c = incremento.
a = multiplicador. m = módulo.

Si $c \neq 0 \rightarrow$ generador mixto
$c = 0 \rightarrow$ generador multiplicativo

Para generar nº pseudoaleatorios distribuidos uniformemente en (0,1):

$$x_i = y_i / m$$

Propiedades

- "m" se relaciona con la longitud de la secuencia y velocidad computacional.
- "a" y "c" se relacionan con la aleatoriedad.
- El menor nº K tal que $y_{n+K} = y_n$ es el período de la secuencia.
- El período de la secuencia está acotado por "m".

Ejemplo: $y_0 = 35$, $a = 13$, $c = 65$, $m = 100$

$y_1 = (13 \cdot 35 + 65) \bmod [100]$

$y_1 = 520 \bmod [100] = 20 \Rightarrow x_1 = \frac{20}{100} = 0{,}2$

$\rightarrow y_2 = (13 \cdot 20 + 65) \cdot \bmod [100]$

$y_2 = 325 \bmod [100]$

$y_2 = 25 \Rightarrow x_2 = \frac{25}{100} = 0{,}25.$

...

③ Generador de Fibonacci - GFSR

Se basa en la ecuación: $x_i = (x_{j-1} \odot x_{i-k}) \bmod [m]$ con $0 < j < K$, donde "$\odot$" es un operador como una suma o XOR (en este caso se llama GFSR).

Hay una versión de GFSR llamada *Mersenne Twister*, que es el algoritmo que usa Python para generar números aleatorios.

ESPACIO DE PROBABILIDAD (Ω, F, P)

* **Espacio muestral** (Ω)

Conjunto de todos los posibles resultados del experimento aleatorio.

Evento → Subconjunto de resultados de un espacio muestral. Puede ser elemental <no se puede dividir, tiro el dado y sale 3> o no elemental <tiro el dado y sale par>

#A : nº de eventos elementales (cardinal).

Los eventos en un espacio muestral satisfacen la teoría de conjuntos:

- $A \cup B = \{w \in \Omega : w \in A \vee w \in B\}$
- $A \cap B = \{w \in \Omega : w \in A \wedge w \in B\}$
- $A - B = \{w \in \Omega : w \in A \wedge w \notin B\}$
- $A^c = \{w \in \Omega : w \notin A\}$

Scanned with CamScanner

* **σ-álgebra de subconjuntos de Ω ($\mathcal{F}$)**

Clase o colección no vacía de eventos (subconjuntos de Ω) cerrada bajo las operaciones de tomar complementos, uniones o intersecciones, es decir, S es σ-algebra si $\{A^c \in S \ \forall A \in S\}$ y $\{\bigcup_i^n A_i \in S, i=1,..,k, A_i \in S\}$

Ej: $\mathcal{F} = \{\phi, \Omega\}, \{\phi, \Omega, A, A^c\}$.

* **Medida de probabilidad (P)**

Función P definida sobre $\mathcal{F}$ que toma valores en el intervalo [0,1], asignando una probabilidad a cada evento en $\mathcal{F}$.

Hay varias formas de definir funciones P:

- Probabilidad subjetiva → Es intuitiva y depende del observador. Ej: si hay 32 equipos en el Mundial, la prob. de ganar es 1/32, pero hay expertos que no creen esto.
- Probabilidad clásica → Dado A un evento: $A \subseteq \Omega \Rightarrow P(A) = \frac{\#A}{\#\Omega}$.
  Ej: la probabilidad de que salga par en un dado:
  $$\frac{\#\{2,4,6\}}{\#\{1,2,3,4,5,6\}} = \frac{3}{6} = \frac{1}{2}$$
- Probabilidad frecuentista → $P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$ con n(A) el número de veces que ocurre el evento A luego de realizar el experimento n veces.
- Probabilidad geométrica → $P(A) = \frac{\text{área}(A)}{\text{área}(\Omega)}$ donde $A \in \Omega$ y $\Omega \in \mathbb{R}^2$ con Ω un espacio muestral con área positiva y finita. Ej: El % de área entre [0,1] que sea menor a 0,5.
- Probabilidad axiomática → Plantea las propiedades que debe satisfacer una función para ser una medida de probabilidad. Ej: postulados de Kolmogorov: una función $P: \Omega \to \mathbb{R}$ es una medida de probabilidad si satisface que
  * $P(\Omega) = 1$ ⟶ certidumbre
  * $P(A) \geq 0$ ⟶ positividad
  * $A \cap B \ \phi \Rightarrow P(A \cup B) = P(A) + P(B)$ ⟶ adición.

CONCEPTOS IMPORTANTES

* **2 eventos independientes**: $(A \perp B)$ si $P(A \cap B) = P(A).P(B)$
  El resultado de uno, no afecta al otro: $P(A|B) = P(A)$
  $P(B|A) = P(B)$

* **Probabilidad condicional**: Es la probabilidad de A dado (que ocurrió) B.
  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ se calcula P(A) en un Ω más reducido, dado que ocurrió B.

* **Regla de la multiplicación**: La probabilidad de que dos eventos A y B ocurran simultáneamente es igual a: $P(B).P(A|B) = P(A).P(B|A)$
  Ejemplo: la probabilidad de que haya ☆ de tipo K con planetas es igual a la prob. de que haya ☆ tipo K por la prob. de que haya planetas dado que hay ☆ tipo K: P(☆K con planetas) = P(☆K).P(planetas|☆K)

Scanned with CamScanner

TEOREMA DE LA PROBABILIDAD TOTAL

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{A_i\}_{i=1}^{\infty}$ una partición de Ω con $P(A_i)>0 \ \forall i$ (A_i abarca todo Ω, son disjuntos) entonces:

(V.A. discreta) $P(B) = \sum_{i=1}^{\infty} P(B|A_i).P(A_i)$ $\forall$ evento $B \in \mathcal{F}$.

(V.A. continua) $P(x) = \int P(x|y).P(y)\,dy$

TEOREMA DE BAYES

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{A_i\}_{i=1}^{\infty}$ una partición de Ω con $P(A_i)>0 \ \forall i$ entonces:

(V.A discreta) $$P(A_i|B) = \frac{P(B|A_i).P(A_i)}{\sum_{k=1}^{\infty} P(B|A_k).P(A_k)} \quad \forall \text{ evento } B \in \mathcal{F}$$

$\rightarrow$ condicional

(V.A. continua) $$P(x) = \frac{P(x|y).P(y) \rightarrow \text{marginal}}{\int P(x|y).P(y)\,dy}$$

Si H=Hipótesis y E=Evidencia obtenida de los datos, entonces H puede ser V o F con alguna probabilidad dada por:

likelihood $$P(H|E) = \frac{P(E|H).P(H) \rightarrow \text{prior o probabilidad anterior}}{P(E) \rightarrow \text{evidencia.}}$$

probabilidad posterior

Si M hace referencia a una familia paramétrica de modelos:

$H = M = $ modelo
$E = D = $ datos

$$P(M, \vec{\theta}|D) = \frac{P(D|M, \vec{\theta}).P(M, \vec{\theta})}{P(D)}$$

FUNCIÓN DE PROBABILIDAD ACUMULADA

prob. en el valor x.

(V.A. continua) $\rightarrow$ Función de densidad de probabilidad $\rightarrow f(x) = P(X=x)$ <PDF>

Función de probabilidad acumulada (distribución) $\rightarrow F(x) = P(X<x)$ <CDF>

prob. de que x caiga en un valor y en los anteriores.

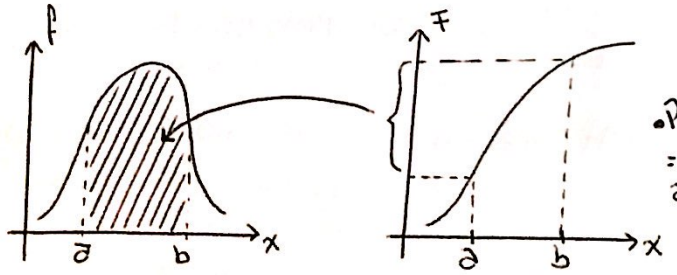
Se relacionan mediante: $$F(x) = \int_{-\infty}^{x} f(x')\,dx'$$

$$F(x): \mathbb{R} \rightarrow [0,1]$$

Propiedades

- $P(x \leq \infty) = 1$ $\quad \lim_{x\to\infty} F(x) = 1$
- $P(x \geq -\infty) = 0$ $\quad \lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to x_0^+} F(x) = F(x_0)$
- $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$ F es siempre creciente.
- $P(a \leq x \leq b) = F(b) - F(a)$

La probabilidad de un evento es el área bajo la curva de la PDF.



- $P(a \leq x \leq b) = \int_a^b f(x)\,dx$

NOTA | Cuando la $f(x)$ es continua no hay que darla sola, sino: $f(x)dx$

Scanned with CamScanner

TEOREMA DE LA PROBABILIDAD TOTAL

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{A_i\}_{i=1}^{\infty}$ una partición de Ω con $P(A_i)>0 \ \forall i$ (A_i abarca todo Ω, son disjuntos) entonces:

(V.A. discreta) $P(B) = \sum_{i=1}^{\infty} P(B|A_i).P(A_i)$ $\forall$ evento $B \in \mathcal{F}$.

(V.A. continua) $P(x) = \int P(x|y).P(y)\,dy$

TEOREMA DE BAYES

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{A_i\}_{i=1}^{\infty}$ una partición de Ω con $P(A_i)>0 \ \forall i$ entonces:

(V.A discreta) $P(A_i|B) = \dfrac{P(B|A_i).P(A_i)}{\sum_{k=1}^{\infty} P(B|A_k).P(A_k)}$ $\forall$ evento $B \in \mathcal{F}$

(V.A. continua) $P(x) = \dfrac{\overbrace{P(x|y)}^{\text{condicional}}.\underbrace{P(y)}_{\text{marginal}}}{\int P(x|y).P(y)\,dy}$

prob. condicional $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ (*)

regla de mult $P(A).P(B|A) = P(B).P(A|B)$

$\Rightarrow P(A \cap B) = P(B).P(A|B) = P(A).P(B|A)$

PROB TOTAL $P(B) = \sum_i P(B|A_i)\,P(A_i)$

(*) $P(A_i|B) = \dfrac{P(B|A_i)\,P(A_i)}{\sum_k P(B|A_k).P(A_k)}$

Si H=Hipótesis y E=Evidencia obtenida de los dato
V o F con alguna probabilidad dada por:

likelihood $P(H|E) = \dfrac{P(E|H).P(H)}{P(E)}$ → prior o prob anterior; $P(E)$ → evidencia.

$P(H|E)$ ↓ probabilidad posterior

Si M hace referencia a una familia paramétrica d

$$P(M,\vec{\theta}|D) = \frac{P(D|M,\vec{\theta}).P(M,\vec{\theta})}{P(D)}$$

FUNCIÓN DE PROBABILIDAD ACUMULADA

(V.A. continua) x

Función de densidad de probabilidad → $f(x) = P(X=x)$ <PDF> (prob. en el valor x.)

Función de probabilidad acumulada (distribución) → $F(x) = P(X<x)$ <CDF> (prob. de que x caiga en un valor y en los anteriores.)

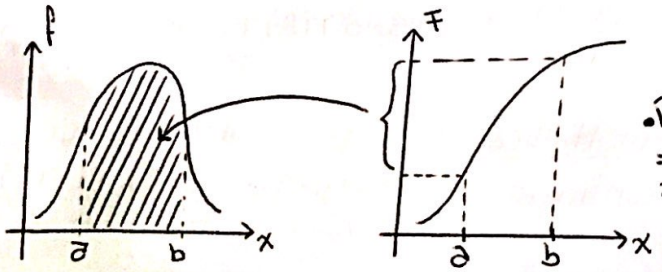
Se relacionan mediante: $F(x) = \int_{-\infty}^{x} f(x')\,dx'$

$F(x): \mathbb{R} \to [0,1]$

Propiedades

- $P(x \leq \infty) = 1$; $\lim_{x\to\infty} F(x) = 1$
- $P(x \geq -\infty) = 0$; $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to x_0^+} F(x) = F(x_0)$
- $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$ F es siempre creciente.
- $P(a \leq x \leq b) = F(b) - F(a)$

La probabilidad de un evento es el área bajo la curva de la PDF.



- $P(a \leq x \leq b) = \int_a^b f(x)\,dx$

NOTA | Cuando la $f(x)$ es continua no hay que darla sola, sino: $f(x)dx$

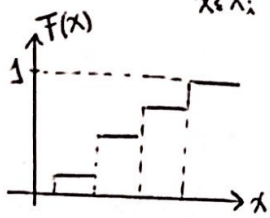
Scanned with CamScanner

(V.A. discreta) Función de masa de probabilidad → $f(x) = P(X=x)$ <PMF>

"forma de escalera" ← Función de probabilidad acumulada → $F(x) = P(X<x)$ <CDF>

Se relacionan mediante: $F(x) = \sum_{x<x_i} f(x=x_i)$

F es continua a trozos:



VARIAS VARIABLES

Si en lugar de tener una V.A se tienen varias, se define el vector aleatorio:

Si $\{X_i\}_{i=1}^{n}$ son V.A. ⇒ $(X_1, \dots, X_n): \Omega \to \mathbb{R}^n$ es un vector aleatorio.

puede ser discreto (de V.A. discretas).
puede ser continuo (de V.A. continuas).

Interpretación geometrica:

- 1 V.A. ⇒ área bajo la curva.
- 2 V.A ⇒ volumen bajo la superficie. } Ej: temperatura y color de ★.

Función de densidad de probabilidad conjunta: $f(x,y) = P(X=x, Y=y)$

Función de probabilidad conjunta acumulada: $F(x,y) = P(X<x, Y\leq y)$ → se deben cumplir ambas simultánea-mente.

$F(x,y) = \mathbb{R}^2 \to [0,1]$

Se relacionan mediante: $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x',y')\,dx'dy'$

Propiedades

- $F(x,y) \xrightarrow[(x,y)\to-\infty]{(x,y)\to\infty} \begin{matrix}1\\0\end{matrix}$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')\,dx'dy' = 1$ siempre normalizada
- $\langle p(x,y)\rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x',y')\cdot f(x',y')\cdot dx'dy'$

Función de densidad marginal

de x: $f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ de y: $f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$

Función de distribución marginal

de x: $F(x) = \lim_{y\to\infty} F(x,y)$ de y: $F(y) = \lim_{x\to\infty} F(x,y)$

Función de densidad condicional

de x dado y: $f(x|y) = \dfrac{f(x,y)}{F(y)}$ de y dado x: $f(y|x) = \dfrac{f(x,y)}{F(x)}$

Variables independientes

Si x e y son independientes (⊥) entonces → $F(x,y) = F(x)\cdot F(y)$
→ $F(x|y) = F(x)$
→ $F(y|x) = F(y)$

Scanned with CamScanner

MOMENTOS

El momento de orden n de una V.A. con función CDF $F(x)$:

(V.A. discreta) $E(x^n) = \langle x^n \rangle = \sum_{i=1}^{n} x_i^n f(x_i)$

(V.A. continua) $E(x^n) = \langle x^n \rangle = \int_{-\infty}^{\infty} x^n dF(x) = \int_{-\infty}^{\infty} x^n \cdot f(x) dx$

MOMENTO DE ORDEN UNO (Esperanza o valor de expectación).

$E(x) = \langle x \rangle = \int_{-\infty}^{\infty} x \cdot f(x) dx$ (V.A. continua) $\| E(x) = \langle x \rangle = \sum_{i=1}^{n} x_i f(x_i)$ (V.A. discreta)

Es un indicador de la ubicación central de la distribución: μ.

[NOTA] El valor de expectación de una función: $E[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot dF(x)$

El valor de expectación de uno: $E[1] = \langle 1 \rangle = \int_{-\infty}^{\infty} dF(x) = \int_{-\infty}^{\infty} f(x) dx = 1$

MOMENTO DE ORDEN DOS (Varianza)

$$\sigma^2 = E[(x-\mu)^2] = E[(x-E(x))^2] = E[x^2] - E[x]^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$$

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot f(x_i) - \mu^2$$

Es un indicador del grado de dispersión de la distribución alrededor de su media.

Desviación estándar: $\sigma = \sqrt{\text{Varianza}} = \sqrt{\sigma^2}$

Propiedades
- $\sigma^2(c) = 0$
- $\sigma^2(x) \geq 0$
- $\sigma^2(c \cdot x) = c^2 \cdot \sigma^2(x)$
- $\sigma^2(c+x) = \sigma^2(x)$

MOMENTO DE ORDEN TRES (Skewness)

Es un indicador de la asimetría de la distribución respecto a su media: $E[x^3]$.

MOMENTO DE ORDEN CUATRO (Kurtosis)

Es un indicador de cuan sobresaliente es el pico: $E[x^4]$.

COVARIANZA

La covarianza de dos V.A, x e y, se define como: $Cov(x,y) = E\left[(x-E(x)) \cdot (y-E(y))\right]$

Propiedad importante $\Rightarrow$ $Cov(x,x) = Var(x) = \sigma^2(x)$.

Coeficiente de correlación: Es la covarianza normalizada a las desviaciones estándares:

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}$$

$Cov(x,y) = 1 \Rightarrow$ correlación total.

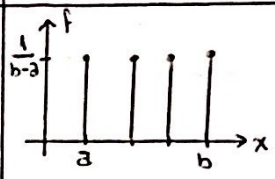
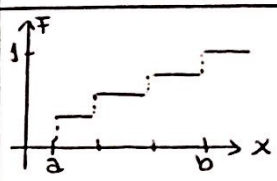
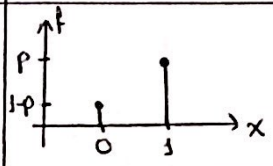
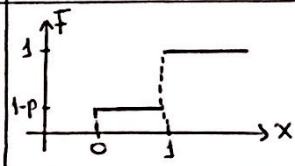
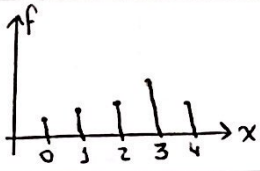
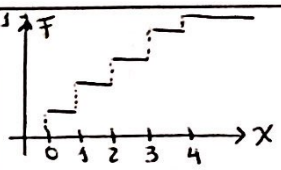
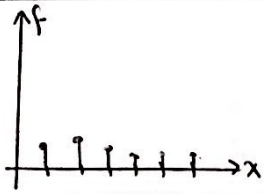
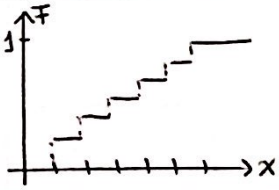
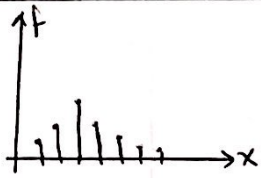
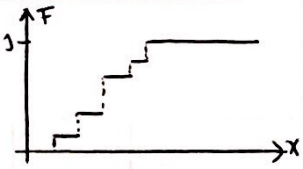
$Cov(x,y) = 0 \Rightarrow$ correlación nula.

$Cov(x,y) = -1 \Rightarrow$ anticorrelación total.

Scanned with CamScanner

DISTRIBUCIONES DE PROBABILIDAD DE LA V.A.

V.A. DISCRETAS

TIPO	MEDIA	VARIANZA	PMF (FUNCIÓN)	PMF (GRÁFICO)	CDF (GRÁFICO)
Uniforme			$f(x)=\begin{cases}\frac{1}{b-a} & x_i=x\in[a,b]\\ 0 & x_i\neq x\in[a,b]\end{cases}$		
Bernulli	p	$p(1-p)=p.q$	$f(x)=\begin{cases}p & x=1\\ 1-p & x=0\end{cases}$ 2 resultados: éxito → con prob. p. fracaso → con prob. 1-p.		
Binomial	np	$np(1-p)=npq$	$f(x)=\binom{n}{x}p^x(1-p)^{n-x}$ n: nº de ensayos p: probabilidad de éxito		
Geométrica	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$f(x)=(1-p)^x.p$		
Poisson	λ	λ	$f(x;\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}$		

Uniforme
son x resultados equiprobables. Ejemplo: dado.

Bernulli
El experimento aleatorio tiene como espacio muestral {éxito, fracaso}. Ejemplo: tirar una moneda y determinar si sale cara. x = nº de éxitos en 1 ensayo.

Binomial
Cuenta el número de éxitos en "n" ensayos de Bernulli independientes entre sí con probabilidad fija p de éxito. x = nº de éxitos en n ensayos.

Si $n\to\infty$ entonces Binomial → Normal.

Si n → chico entonces Binomial → Poisson.

Ejemplo: número de caras en cinco lanzamientos de una moneda.

Geométrica
Cuenta el número de intentos x que se requieren para observar un único éxito en una sucesión ∞ de ensayos independientes de Bernulli con prob. de éxito p.

Ejemplo: el número de veces que se debe lanzar un dado para que se observe un seis.

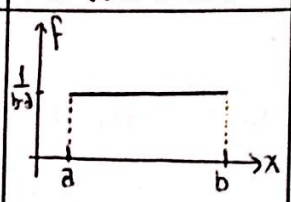
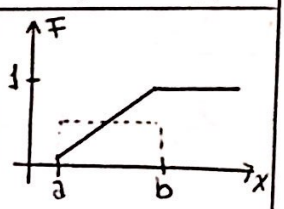
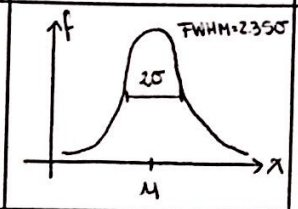
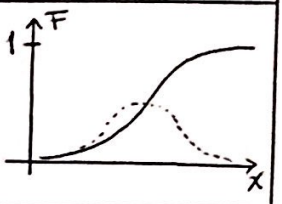
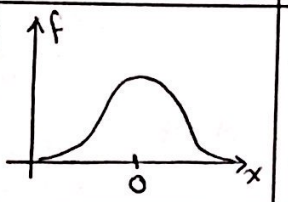
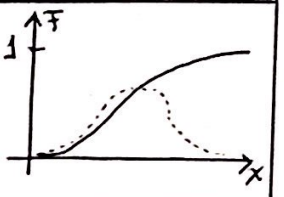
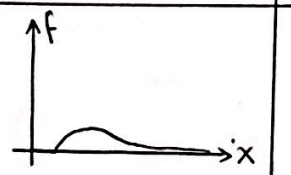
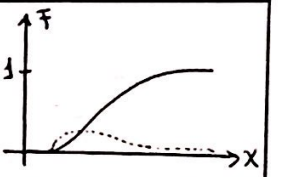
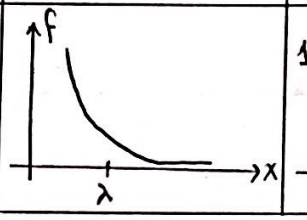
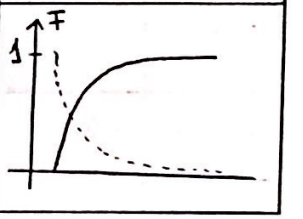
Poisson
Expresa la probabilidad que ocurra un determinado número de eventos durante cierto período de tiempo o espacio. Ejemplo: goles en un partido de futbol.

λ = nº de veces que se espera que ocurra el fenómeno por unidad de tiempo o espacio.

x = nº de veces que ocurre éxito en ese mismo tiempo o espacio.

Scanned with CamScanner

VA CONTINUAS

TIPO	MEDIA	VARIANZA	PDF (FUNCIÓN)	PDF (GRÁFICO)	CDF (GRÁFICO)
Uniforme	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$f(x;a,b)=\begin{cases}\frac{1}{b-a} & x\in[a,b]\\ 0 & x\notin[a,b]\end{cases}$		
Gaussiana	μ	σ^2	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$		
T-Student	0	$\frac{k}{k-2}$	$\frac{Z}{\sqrt{U/k}}$, $Z\sim N(0,1)$, $U\sim\chi_k$		
Chi cuadrado	k	2k	$\sum_{t=1}^{k} Z_t^2$, $Z_t \overset{iid}{\sim} N(0,1)$		
Exponenual	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x} & x\geq 0\\ 0 & x<0\end{cases}$		

Uniforme

Todo el intervalo es igualmente probable. Ejemplo: modelar fechas de nacimiento, ya que todos los momentos del año son igualmente probables.

Gaussiana o Normal

Función de densidad simétrica, con máximo en $x=\mu$ y dos puntos de inflexión en $x=\mu\pm\sigma$. Se usa para representar variables aleatorias de valor real que se supone que se producen de forma aditiva por muchos efectos pequeños. Ejemplo: modelar la altura de las personas, resultado de muchos factores genéticos y ambientales pequeños.

T-Student

Surge al estimar la media de una población distribuida normalmente en situaciones en las que el tamaño de la muestra es pequeño y se desconoce la desviación estándar de la población.

Chi-cuadrado

Es la suma de k variables aleatorias (k grados de libertad) normales estándar cuadráticas independientes e idénticamente distribuidas. Ejemplo: pruebas de hipótesis y construcción de intervalos de confianza.

Exponencial

Es el análogo continuo de la distribución geométrica. Ejemplo: modelar tiempos de espera.

Scanned with CamScanner

TEOREMAS IMPORTANTES

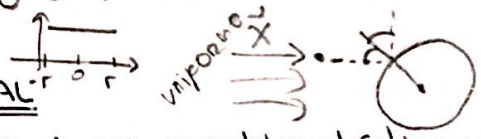
CAMBIO DE VARIABLES ALEATORIAS

Idea → tengo V.A con dist. conocida $\xrightarrow{\text{transf}}$ obtenes nuevo conj. V.A. con dist. desc.

$$\underset{\substack{f(x)\\ \text{conocida}}}{\vec{x}} \longrightarrow \boxed{T} \longrightarrow \underset{\substack{\tilde{f}(x)\\ \text{desc.}}}{\vec{y}}$$

La $T + \vec{x}$ y $f(x)$ → dan $\tilde{f}(x)$.

Empiezo con $\vec{y}$ y quiero conocer $\tilde{f}$ → invento $\vec{x}$ con f conocida para hallar $\tilde{f}$



$\vec{y}$ = desparrame de las hipótesis → choque elástico

TEOREMA DEL LÍMITE CENTRAL

Sea $\{x_i\}_{i=1}$ un conjunto de "n" variables idénticamente independientes con $F(x)$ arbitraria, con medias $\{\mu_i\}_{i=1}$ y varianzas $\{\sigma_i^2\}_{i=1}$ finitas. Sean, además $S_n = \sum_{i=1} x_i$, $\mu = \sum_{i=1} \mu_i$ y $\sigma^2 = \sum_{i=1} \sigma_i^2$.

$\Rightarrow S_n$ tiende en probabilidad a $N(\mu, \sigma^2)$

Demostración

Calculamos: $\langle e^{ik(x_N-\mu)} \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{ik(x_N-\mu)} \cdot F(\vec{x})\, dx$

$$= \langle e^{ik(x_1+x_2+\dots+x_N-N\mu)/N} \rangle$$

$$= \langle e^{ik[(x_1-\mu)+\dots+(x_N-\mu)]/N} \rangle$$

iid ⇒ siguen la misma $F(x)$

$$= \langle e^{ik(x-\mu)/N} \rangle^N$$

promedio de cada término

$$= \langle 1 + \frac{ik}{N}(x-\mu) - \frac{k^2}{2N^2}(x-\mu)^2 + \mathcal{O}\left(\frac{k^3}{N^3}\right) \rangle^N$$

$$= \left[1 + 0 - \frac{k^2}{2N^2}\sigma^2 + \mathcal{O}(3)\right]^N$$

$$\lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N = e^x$$

$$= e^{-k^2\sigma^2/2N}$$

Entonces tengo $\langle e^{ik(x_N-\mu)} \rangle = e^{-k^2\sigma^2/2N}$ y si hago la antitransformada:

$$F(x_N-\mu) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-k^2\sigma^2/2N} \cdot e^{-ik(x_N-\mu)}\, dk \underset{N\to\infty}{=} \sqrt{\frac{N}{2\pi\sigma^2}} \cdot e^{-N(x_N-\mu)^2/2\sigma^2}$$

Sigue una distribución normal.

LEY DE LOS GRANDES NÚMEROS

Sea $\{x_i\}_{i=1}^n$ un conjunto de "n" variables aleatorias independientes e idénticamente distribuidas con media μ y sea $S_n = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ entonces:

$$\lim_{n\to\infty} S_n = \mu$$

En criollo ⇒ la media muestral da una buena aproximación de la media poblacional μ cuando n es grande, sin importar la distribución original

Scanned with CamScanner

FUNCIONES GENERADORAS

FUNCIÓN GENERADORA DE PROBABILIDAD (FGP)

$$G_x(t) = E[t^x] = \langle t^x \rangle = \int_{-\infty}^{\infty} t^x \cdot f(x)\, dx \quad \text{(V.A. continua).}$$

$$\langle t^k \rangle = \sum_{k=0}^{\infty} t^k \cdot \underbrace{f(k)}_{f(k)=P(X=k)} \quad \text{(V.A. discreta).}$$

A partir de la FGP se puede construir la $f(k) = \left.\dfrac{\partial^k G(t)}{\partial t^k}\right|_{t=0}$

Si x, y son V.A. independientes $\Rightarrow G_{x+y}(t) = G_x(t) + G_y(t)$.

FUNCIÓN GENERADORA DE MOMENTOS (FGM)

$$M_x(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) \cdot dx \quad \text{(V.A. continua).}$$

$$E[e^{tk}] = \sum_k e^{tk} \cdot f(k) \quad \text{(V.A. discreta).}$$

Cuando existe la FGM determina univocamente la PDF (probability density funtion) $\Rightarrow$ La PDF sería la transformada de Fourier inversa de la FGM.

$$M_x(t) = G_x(e^t)$$

FUNCIÓN GENERADORA DE CUMULANTES (FGC)

$$g_x(t) = \text{Log}\,[E(e^{tx})] = \sum_{n=1}^{\infty} k_n \frac{t^n}{n!} = \text{Log}(FGM) \rightarrow g_x(t) \text{ es serie de potencias con coeficientes } k_n = \text{cumulantes de } F.$$

FUNCIÓN CARACTERÍSTICA DE UNA V.A. X (FC)

$$\phi_x(t) = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} \cdot f(x)\, dx = \langle e^{ikx} \rangle$$

Si su parte real e imaginaria existe y es finita $\Rightarrow \underbrace{\phi_x(t) = M_x(it) = G_x(e^{it})}_{\text{relación de la FC con la FGM y la FGP.}}$

Propiedad reproductiva de la distribución normal

Sea $y = \sum_{i=1}^{n} x_i$ con $x_i \sim N(\mu, \sigma^2)$ iid, entonces y se distribuye como $N(n\mu, \sqrt{n}\sigma^2)$. Es decir que la suma de V.A normales con la misma media μ y varianza σ^2, es una variable con distribución normal con media $n\mu$ y varianza $n\sigma^2$.

Proposición: Si $y = \frac{1}{n}\sum_{i=1}^{n} x_i$ con $x_i \sim N(\mu, \sigma^2)$ <promedio de V.A. normales> iid, entonces $y \sim N(\mu, \frac{\sigma^2}{n})$.

$$\left.\begin{array}{l} * \; y = \sum_{i=1}^{n} x_i \text{ con } x_i \sim Bin(n_i, p) \Rightarrow y \sim Bin\left(\sum_{i=1}^{n} n_i, p\right) \\ * \; y = \sum_{i=1}^{n} x_i \text{ con } x_i \sim Poisson(\lambda_i) \Rightarrow y \sim Poisson\left(\sum_{i=1}^{n} \lambda_i\right) \end{array}\right\} \text{propiedad reproductiva.}$$

Scanned with CamScanner

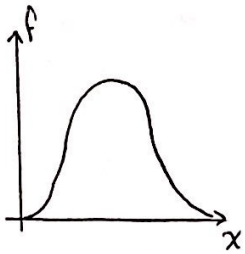
MÉTODO DE LA FUNCIÓN INVERSA

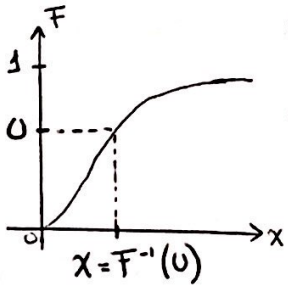
Lo que hace este método es generar un conjunto de números aleatorios que siguen una función de desidad/masa de probabilidad dada.

Sea φ una función de probabilidad acumulada INVERTIBLE y sea U una V.A. uniforme en $(0,1)$, entonces la V.A $x = \varphi^{-1}(U)$ tiene distribución de probabilidad acumulada $F(x) = \varphi$.

(V.A. continua)

Quiero un sorteo de no aleatorios con distribución f.



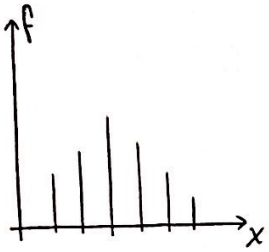


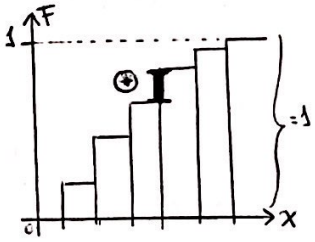
① Sorteo una variable aleatoria uniforme en $(0,1) \to U$

② Le aplico la función inversa de F y me devuelve $x = F^{-1}(U)$

③ Repito y obtengo un sorteo de la variable x con función de distribución acumulada $F(x)$.

(V.A. discreta)





① Sorteo $U \sim U(0,1)$ y cae en algún lado.

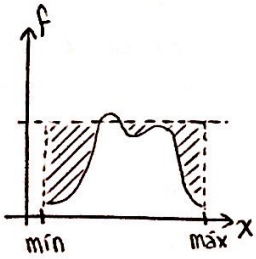
② Si la barra es más larga, es más probable que U caiga ahí.

③ Veo a que pared llega y asigno a ese x_i.

Ejemplo: Experimento del Buffón.

Consiste en dejar caer una aguja sobre una hoja rayada y anotar las veces que la aguja cae sobre alguna de las rayas.

MÉTODO DE ACEPTACIÓN-RECHAZO



Tengo una distribución con una forma "extraña" y quiero sortear valores de x que sigan dicha distribución f. Entonces, tomo una distribución de referencia (una uniforme), sorteo un x en la distribución $U(\text{mín}, \text{máx})$ y me quedo con los que caigan dentro de la distribución f.

Es decir, el algoritmo obtiene una muestra pseudoaleatoria proveniente de una V.A x con función de densidad f utilizando una muestra pseudoaleatoria de una V.A. Y con función de densidad d:

① Genera Y con función de densidad d.

② Genera $U \sim U(0,1)$ independiente de Y.

③ Si $U \leq \frac{f(\cdot)}{g(Y)}$ entonces $X = Y$, de lo contrario repetir ①.

Con M iteraciones se consigue una muestra pseudoaleatoria.

Scanned with CamScanner

NOTA La parte rayada suele ser mayor que la distribución f y por eso es poco eficiente => la solución es usar otra distribución de referencia más "pegada" a la que se quiere reproducir.

MÉTODO BOX MÜLLER

Los dos métodos anteriores no se usan para generar muestras pseudoaleatorias con función de densidad de distribución gaussiana. El método de Box Müller genera números pseudoaleatorios gaussianos. Esto se debe a que la función de distribución acumulada de la gaussiana no es invertible (no anda método de la función inversa) ni tampoco se puede encuadrar la misma gaussiana, $x \in (-\infty, \infty)$, (no anda método de aceptación-rechazo).

Se toman dos variables distribuidas uniformemente y se las transforman a dos V.A. independientes con una distribución normal estándar.

Digamos que U_1 y U_2 son V.A. con dist. uniforme en el intervalo $(0,1)$. La transformación de Box Müller crea nuevos Z_0 y Z_1 con distribución normal estándar: la derivación está basada en que podemos representar cualquier punto en el plano cartesiano x, y a través de coordenadas polares, con un radio R y un ángulo θ.

$$\left.\begin{aligned} Z_0 &= R \cdot \cos\theta = \sqrt{-2\ln(U_1)} \cdot \cos(2\pi U_2) \\ Z_1 &= R \cdot \operatorname{sen}\theta = \sqrt{-2\ln(U_1)} \cdot \operatorname{sen}(2\pi U_2) \end{aligned}\right\} \quad \begin{aligned} R &= \sqrt{-2\ln(U_1)} \\ \theta &= 2\pi U_2 \end{aligned}$$

Entonces, se sortean números con distribuciones:

$$R^2 \sim \exp\left(\tfrac{1}{2}\right) \quad y \quad \theta \sim U(0, 2\pi)$$

Se transforman las variables $Z_0 = R\cos\theta$ y $Z_1 = R \cdot \operatorname{sen}\theta$ y se consiguen 2. V.A independientes con distribución normal estándar.

MÉTODOS MONTECARLO

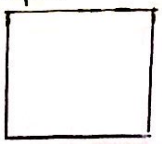
leer cuando llegue a pág 16

Los métodos MonteCarlo son una familia de métodos usados para generar números aleatorios a partir de pruebas aleatorias repetidas.

Supongamos que se quiere conocer la función distribución que sigue una muestra, entonces se debe conseguir una receta de manera que los random caigan con alguna probabilidad.

La diferencia con otros métodos (como los anteriores mencionados) es que en los métodos MonteCarlo se tiene información de algunas características, pero no se tiene la CDF o la PDF como en los otros métodos.

Ejemplo:



Para conocer la razón entre el área del círculo y el área del cuadrado, se tiran randoms dentro del cuadrado y se cuentan cuantos caen dentro y fuera del círculo:

$$\# = \frac{\pi R^2}{4R^2} = \frac{\bigcirc}{\square}$$

Este es un método MonteCarlo.

Scanned with CamScanner

CAMINATAS ALEATORIAS

Un salto permite desplazarse en x e y tomando valores en el rango $[a,b]$. Desde una función uniforme $U \sim U(0,1)$ se obtiene un número random entre 0 y 1 y se reescala de manera que:

$$\left.\begin{array}{l} x \to U(0,1) \\ x' \to U(a,b) \end{array}\right\} x' = x.(b-a)+a$$

Una caminata aleatoria queda determinada por un conjunto o sucesión de saltos.

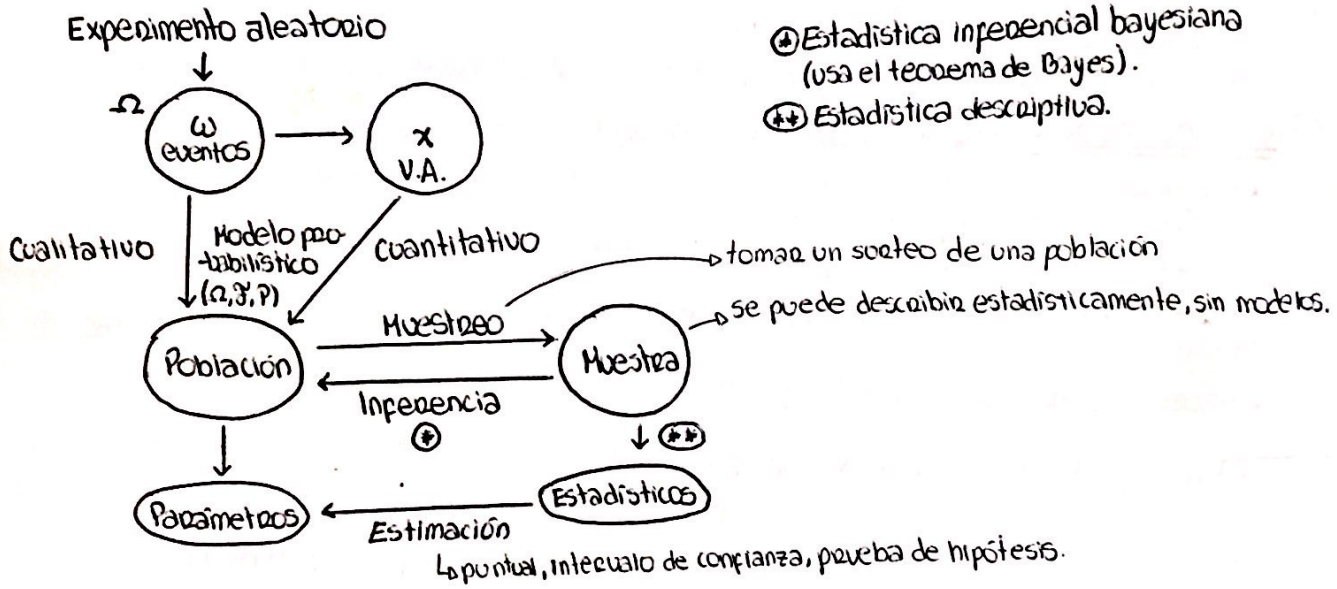
CADENAS DE MARKOV

Una cadena de Markov es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior. En efecto, las cadenas tienen "memoria" y esto condiciona los eventos futuros.

ALGORITMO DE METRÓPOLIS-HASTINGS

Es un método MCMC (Monte Carlo por cadena de Markov, se basan en el camino de las cadenas de Markov cuyas leyes estacionarias son las distribuciones a muestrear) que usa caminatas aleatorias con el objetivo de obtener un muestreo aleatorio de una distribución de probabilidad cuando el muestreo directo es difícil.

ESTADÍSTICA Y PROBABILIDAD



TEORÍA DE PROBABILIDAD

Describe, mediante un modelo matemático, los fenómenos aleatorios.

INFERENCIA ESTADÍSTICA

Infiere características de la población a partir del estudio de una muestra de dicha población. En general, se elige una distribución teórica, se estiman parámetros, se hacen test de bondad de ajuste y se simula a partir de la distribución teórica.

ESTADÍSTICA DESCRIPTIVA

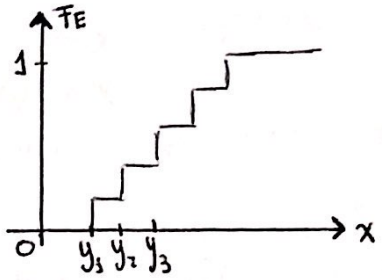
Son herramientas para describir un conjunto de datos como histogramas o función acumulada empírica (ECDF).

Scanned with CamScanner

ECDF

Se tienen datos observados $y_1, \dots, y_n$ (un conjunto de V.A) idénticamente distribuidos, la distribución acumulada empírica (ECDF) sale de ordenar de forma creciente los y_i:

$$F_E(x) = \begin{cases} 0 & x < y_1 \\ 1/n & y_1 < x < y_2 \\ \vdots & \vdots \\ i/n & y_i < x < y_{i+1} \\ 1 & y_n \leq x \end{cases}$$



Se puede hacer con datos discretos o continuos (agruparlos en bines y hacer la ECDF del histograma).

También se puede hacer una interpolación entre los escalones (una recta) queda más suave.

SISTEMAS REALES

Los sistemas reales tienen puentes de aleatoriedad, para estudiar sistemas reales se debe representar cada puente de aleatoriedad con una PDF ¿Como elegir las distribuciones?.

① USANDO LOS DATOS

Se usa cuando no se tiene un modelo teórico para usar.

② USANDO LA DISTRIBUCIÓN EMPÍRICA

Se usa cuando no se pueden ajustar los datos a una distribución teórica.

③ USANDO TÉCNICAS DE INFERENCIA ESTADÍSTICA

* Elegir una distribución en base a los datos observados
  - Conocer el origen de los datos.
  - Estimar parámetros a partir de los datos.
  - Hacer histogramas, q-q plots, p-p plots
* Estimar los parámetros de la distribución
* Hacer pruebas de bondad de ajuste.

NOTA 1 Q-Q plots es un método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación. Q proviene de cuartiles.

NOTA 2 P-P plots o gráfico de valor p sirve para evaluar qué tan cerca coinciden dos conjuntos de datos, o para evaluar qué tan cerca se ajusta un conjunto de datos a un modelo en particular.

NOTA 3

PARÁMETRO	ESTIMADOR	ESTIMA
Min, Max	$X_{min}, X_{máx}$	Rango
Media μ	Promedio $\bar{x}$	Tendencia central
Mediana	$\hat{m}$	"
Varianza σ^2	s^2	Variabilidad, ancho

Scanned with CamScanner

MAPA CONCEPTUAL

DATOS + MODELOS → RESULTADOS — INFERENCIA

DATOS + RESULTADOS → MODELOS — MACHINE LEARNING, REDES.

- TÉCNICAS DE MUESTREO → Es lo que se usa para construir la M.A. con la que se hace inferencia de parámetros.
- ESTIMACIÓN DE PARÁMETROS — mediante →
  - Estimación puntual.
  - Intervalos de confianza.
  - Pruebas de hipótesis.
    - Paramétricas.
    - No paramétricas.
  - métodos de construcción del estimador
    - maximum likelihood.
    - método de momentos.

  } (lo que se construye es la fórmula del estimador).

TÉCNICAS DE MUESTREO

El muestreo estudia la relación entre la distribución de un parámetro (población) y la distribución de su estimador (muestra).

VENTAJAS DE USAR MUESTRAS
- Accesibilidad (ya que a veces la población es inaccesible en su totalidad).
- Rapidez (el analisis se da sobre "menos" datos).
- Menor costo (").

TIPOS DE ELECCIÓN DE LA MUESTRA

- **ALEATORIO**: Todos los valores tienen la misma probabilidad de ser seleccionados.
- **POR CONGLOMERADOS**: Divide la población en subgrupos y selecciona aleatoriamente de cada subgrupos. Ej: niños, adultos, ancianos.
- **ESTRATIFICADO**: Divide la población en subgrupos y selecciona una M.A. de c/u, pero de manera que los subgrupos en la muestra total mantengan la misma proporción que la población.
- **CONVENIENTE**: Se incluyen los valores mas accesibles y disponibles.
- **SISTEMÁTICO**: Muestrea elementos de la población a intervalos regulares. Ej: tomar una lámpara cada 10 casas para probar que funcione.
- **MULTISTAGE**: 1º muestreo por conglomerados y el 2º aleatorios.
- **SNOWBALL**: Elige datos por recomendación de otros datos.

Scanned with CamScanner

ESTIMACIÓN DE PARÁMETROS

¿QUÉ ES UN ESTIMADOR? Es una función (fórmula) que "come" una M.A. y "escupe" un valor que estima un parámetro. Ej: $\underset{\text{parámetros}}{\theta} = \mu, \sigma^2 \rightarrow \underset{\text{estimadores}}{\hat{\theta}} = \bar{x}, s^2$.

MÉTODOS DE CONSTRUCCIÓN DE LOS ESTIMADORES

MÁXIMO LIKELIHOOD

Recordando Bayes:

$$\underbrace{P(M,\theta|D)}_{\text{posterior}} = \frac{\overbrace{P(D|M,\theta)}^{\text{likelihood}} \cdot \overbrace{P(M,\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

La función $\hat{\theta}$ se obtiene maximizando $P(M,\theta|D)$. Pero teniendo en cuenta que:

Prior → $P(M,\theta)$ es lo que ya sabemos del modelo.
→ Es cte para cada experimento.

Evidencia → $P(D)$ es cte si tengo siempre el mismo conjunto de datos.
→ Cambia si cambio los datos.

Lo que queda:

$$P(M,\theta|D) \propto P(D|M,\theta) = \mathcal{L}$$

Los parámetros que maximizan la probabilidad a posterior también maximizan el likelihood.

Sea una M.A. $\{x_i\}_{i=1}^{n}$ iid (los datos), buscamos los mejores parametros que describen los datos (describir mejor ≡ mayor probabilidad posterior). Si la muestra tiene PDF $f(x)$, entonces, dado que son independientes e idénticamente distribuidos:

$$\mathcal{L} = P(D|M,\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta) = \prod_i f(x_i|\theta)$$

Luego, $\boxed{\mathcal{L} = \prod_{i=1}^{n} f(x_i|\theta)}$ es la <u>función de maximum likelihood</u> y el valor θ en donde $\mathcal{L}$ alcanza su máximo es el <u>estimador de maximum likelihood</u>.

Por conveniencia, como log es una función continuamente creciente, maximizar la probabilidad del $\mathcal{L}$ es lo mismo que maximizar la probabilidad del $\log(\mathcal{L})$:

$$\log(\mathcal{L}) = \sum_{i=1}^{n} \log\left(f(x_i|\theta)\right)$$

Esto es porque:

$$\log(A \cdot B) = \log A + \log B$$

Scanned with CamScanner

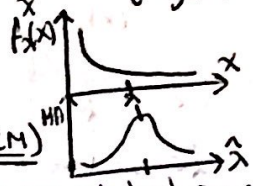
MÉTODO DE MOMENTOS

La función $\hat{\theta}$ se obtiene de igualar los momentos observados (de la muestra) y los teóricos (de la población). Se usan tantos momentos como parámetros a estimar.

Ejemplo

Distribución exponencial $\rightarrow$ $\bar{x} = \frac{1}{\lambda} = E(x) \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$

$$\left.\begin{array}{l} f_\lambda(x) \rightarrow E(x) = \frac{1}{\lambda} \\ MA\{x_i\}_{i=1}^{n} \rightarrow \bar{x} \end{array}\right\} \hat{\lambda} = \frac{1}{\bar{x}}$$



$\hat{\theta}$: Estimador

θ: Parámetro

CALIDAD DEL ESTIMADOR

ERROR CUADRÁTICO MEDIO (ECM)

El ECM es una propiedad del estimador que evalúa su calidad en cuanto a su variación (var) y sesgo (bias, $E(\hat{\theta}) = \theta$).

$$ECM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] = \underbrace{Var(\hat{\theta})}_{\text{varianza}} + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{sesgo}}$$

Si el $\hat{\theta}$ es insesgado $\Rightarrow$ ECM = Var.

Ejemplo: Media muestral ($\bar{x}$)

- Estimador insesgado ($E(\bar{x}) = \mu$)
- Consistente (ley grandes nº: $\lim_{n \to \infty} \bar{x} = \mu$)
- Eficiente de μ.
- $ECM(\bar{x}, \mu) = Var(\bar{x}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(x_i) = \frac{\sigma^2}{n}$

Los estimadores de una M.A. son V.A. Si mido $\bar{x}$ en una M.A una sola vez no aseguro encontrar la media, pero si la muestra no está sesgada, la dist. de densidad de prob. de $\bar{x}$ estará centrada en la media ($E(\bar{x}) = \mu$) y tendrá varianza $\frac{\sigma^2}{n}$ con n el nº de la M.A: $N(\mu, \frac{\sigma^2}{n})$.

Ejemplo 2: Varianza muestral (S^2)

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ • Estimador insesgado.

Al proponer como estimador $\xi = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ pensando en el segundo momento, se llega a que $E(\xi) = \frac{n-1}{n} \sigma^2$, lo que implica que ξ es sesgado (para que sea insesgado $E(\xi) = \sigma^2$). Entonces se propone como estimador insesgado de σ^2 a $S^2 = \frac{n}{n-1} \cdot \xi$.

TÉCNICAS DE ESTIMACIÓN DE PARÁMETROS

ESTIMACIÓN PUNTUAL

Se hace una estimación numérica de un parámetro en una M.A. Es literalmente evaluar la función $\hat{\theta}$ usando una M.A.

Se busca que los estimadores sean:

- INSESGADOS $\Rightarrow E(\hat{\theta}) = \theta$
- CONSISTENTES $\Rightarrow$ Al aumentar el tamaño de la M.A. se aproxima al parámetro.
- EFICIENTES $\Rightarrow$ Es el estimador de menor varianza.
- SUFICIENTES $\Rightarrow$ Utiliza toda la información obtenida de la muestra.

Scanned with CamScanner

INTERVALOS DE CONFIANZA (IC)

Un estimador por intervalo de confianza es un intervalo aleatorio (L, R) que contiene al valor verdadero del parámetro con cierta probabilidad dada por el "nivel de confianza" $1-\alpha$. Es decir, dada una M.A. $\{X_i\}_{i=1}^{n}$ se busca que:

$$P\left[L(X_i) \leq \theta \leq R(X_i)\right] = 1-\alpha$$

Ⓐ IC PARA MODELOS PROBABILÍSTICOS

Se propone un modelo y a partir de él se calcula el I.C. Dada la PDF de un conjunto de datos, una M.A, el intervalo de confianza cumple:

$$\int_L^R f(\theta)\,d\theta = 1-\alpha$$

donde α: significancia. Entre más chico sea α, el intervalo es más pequeño y, por lo tanto, más exigente.

Nivel de confianza del 95% implica $\alpha = 0{,}05$ y del 99% implica $\alpha = 0{,}01$.

NOTA 1 ¿Qué es un cuantil?

Son puntos que dividen la PDF de una M.A. en intervalos regulares. Se denotan con un número entre 0 y 1 → Indica la fracción de los valores de la M.A. que son menores que ese valor:

$$\int_{-\infty}^{\square} f(x)\,dx = \text{cuantil}$$

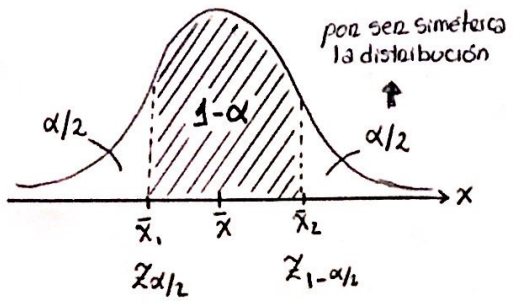
NOTA 2 $\alpha/2$ → cuantil de orden 0,025

$1-\alpha/2$ → cuantil de orden 0975

* IC PARA LA MEDIA CON VARIANZA CONOCIDA

Quiero estimar μ teniendo σ como dato. Esto se puede usar cuando la M.A. de datos tienen una distribución normal (por lo tanto, $\bar{x}$ también por propiedad reproductiva) o cuando la M.A. tiene distribución desconocida pero $n > 30$ (por lo tanto, $\bar{x}$ tiene dist. normal por el TLC).

Normalizando: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z \Rightarrow N(0,1)$ Z tiene distribución normal estándar.



$$P(\bar{x}-\mu \leq \mu \leq \bar{x}+\mu) = 1-\alpha$$

$$\Rightarrow P\left(\bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Este intervalo contiene a $\bar{x}$ con un $100 \cdot (1-\alpha)\%$ de confianza:

$$\boxed{IC = \left[\bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]}$$

$1-\alpha$: nivel de confianza = prob. de que el intervalo contenga al estimador.

IDEAL ⇒ estrechos con alto nivel de confianza.

Scanned with CamScanner

* IC PARA LA MEDIA CON VARIANZA DESCONOCIDA

Podemos dividir en tres casos:

- **CASO 1**: No se conoce la distribución de la población y $n<30$. No es posible predecir la distribución de la media muestral.
- **CASO 2**: La población se distribuye normalmente o no, pero $n \geq 30$, por lo que la media muestral se distribuye normalmente por el TLC.

En este caso, el estadístico es $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ ya que s^2 (varianza muestral) sí es buen estimador de σ^2.

El IC queda:

$$\boxed{IC = \left[\bar{x} + Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} ; \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]}$$

- **CASO 3**: La población se distribuye normalmente y $n<30$. La media muestral sigue la distribución T-Student.

Construimos el estimador t que tiene distribución T-Student:

$$\overset{n-1}{gl} \leftarrow t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \cdot \frac{\sigma}{\sigma} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \cdot \frac{1}{s/\sigma} = \frac{Z}{s/\sigma} \Rightarrow t = \frac{Z}{s/\sigma} \left\} \begin{array}{l} Z \sim N(0,1) \\ \chi^2 \text{ con } n-1 \text{ g. libert.} \end{array} \right.$$

La distribución T-Student es un "cociente" entre la distribución normal estándar y la dist. chi-cuadrado.

¿Por qué aparece chi-cuadrado?

Si $\{X_i\}_{i=1}^{n}$ es una M.A. con distribución $N(0,1)$ entonces $\chi_n^2 = \sum X_i^2$ se distribuye como chi-cuadrado con "$n-1$" grados de libertad: ↦ restricción.

$$\cancel{\chi_n^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{n-1}{\sigma^2}S^2 + \chi_1^2 \Rightarrow \frac{S^2}{\sigma^2} = \frac{\chi_n^2 - \chi_1^2}{n-1} = \frac{\chi_{n-1}^2}{n-1}}$$

~~Luego, S/σ sigue una dist. χ^2 con $n-1$ grados de libertad.~~

gl: variables libres - restricciones.

La T-Student se distribuye simétricamente alrededor de $\bar{x}=0$ y solo depende de los grados de libertad dados por el tamaño de la muestra.

Si la muestra es más chica entonces la dist. es más ancha $\Rightarrow S$ es más grande.

Si la muestra es más grande entonces $S \to \sigma$ y $t \to Z \sim N(0,1)$.

El IC queda:

$$\boxed{IC = \left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} ; \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]}$$

En este intervalo, $\bar{x}$ se encuentra con un $100 \cdot (1-\alpha)\%$ de confianza.

Scanned with CamScanner

* IC PARA PROPORCIONES

Sean $\{X_i\}_{i=1}^{n}$ V.A. de Bernoulli con probabilidad p de éxito. Considerando que "n" es grande:

$$Bin(n,p) \sim N(np, np(1-p))$$

Se puede estimar el parametro p con $\hat{p} = \bar{X}$ y el parámetro Var(p) con

$$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{\bar{x}(1-\bar{x})}{n}$$

Tiene distribución normal

Así, el IC queda:

$$\boxed{IC = \left[\bar{x} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} ; \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]}$$

donde el intervalo contiene a $\hat{p} = \bar{x}$ con un $100.(1-\alpha)\%$ de confianza

(B) IC CON BOOTSTRAPPING

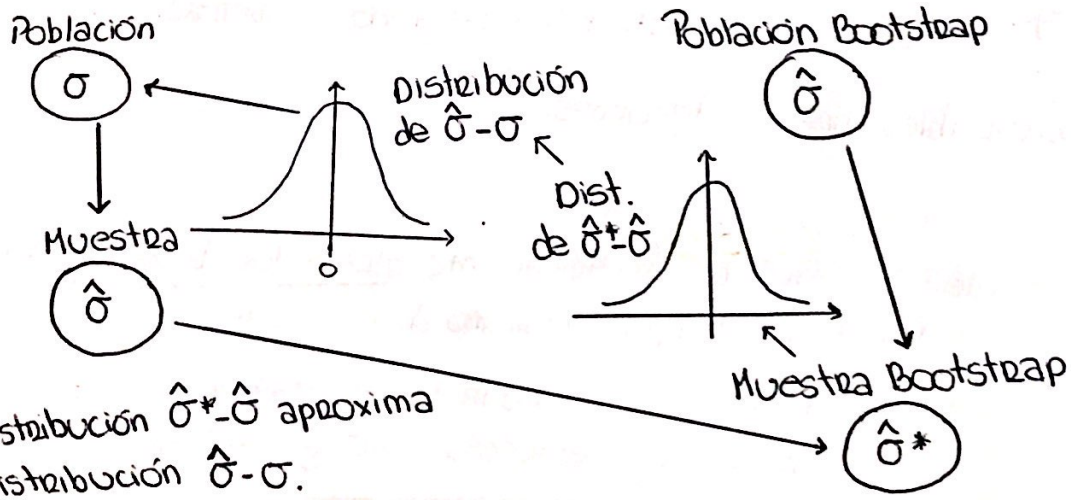
Esto se usa cuando los datos distribuidos normalmente, no se tienen muestras grandes o no se conoce la distribución de la población.

Se usa el método bootstrapping. ¿Qué es el bootstrapping?

Es una técnica de remuestreo que permite estimar el sesgo o la varianza (a partir del ECM) y construir un intervalo de confianza. El procedimiento es:

- Dada una MA. con "n" observaciones, se la trata como si fuera la población completa y se extraen B muestras
- Para cada muestra B se estima el parámetro de interés y se observa la distribución dada por las muestras Bootstrap.

"La muestra es a la población lo que la muestra Bootstrap es a la muestra".



La distribución $\hat{\sigma}^*-\hat{\sigma}$ aproxima la distribución $\hat{\sigma}-\sigma$.

y esto lo hace con una incerteza dada por el ECM.

NOTA Cuando el estimador es insesgado el ECM es el estimador de la varianza.

Scanned with CamScanner

Dada $\{x_i\}_{i=1}^{n}$ MA. con distribución $F(x,\theta)$ se busca $\theta = \theta(F)$, entonces

$\hat{\theta} = g(x_1, \dots, x_n)$ es una función a partir de las observaciones.

La incerteza la busco en el ECM:

$$\underbrace{ECM_e}_{\text{aproximación bootstrap al error cuadrático medio}}(\hat{\theta}, \theta) = \sum_{\text{todas las permutaciones}} (g(\vec{x}) - \theta(F_e))^2 \cdot \frac{1}{n^n}$$

Esto sale del hecho de que no se tiene la F pero se la aproxima por la Fe y los parámetros también.

Pero como es muy costoso la suma de n^n términos, se hace una aproximación Montecarlo (se sortean de forma random B términos) y de esta forma:

$$ECM_e(\hat{\theta}) \simeq \frac{\sum_{j=1}^{B} y_j}{B}$$

PRUEBAS DE HIPÓTESIS

A veces, la inferencia no se plantea en términos del valor de un parámetro (estimación puntual) o de sus barras de error (estimación por IC), sino por medio de la verificación o falsificación de una proposición.

Una prueba de hipótesis es un procedimiento para decidir si se acepta o se rechaza una hipótesis:

1) Formular el problema definiendo el parámetro que se quiere estimar y su estimador.
2) Establecer H_0 y H_a.
3) Definir el nivel de significancia α.
4) Calcular el estadístico de prueba.
5) Calcular la zona/región de aceptación y la de rechazo.
6) Validar las hipótesis.

* HIPÓTESIS ESTADÍSTICAS

- **HIPÓTESIS NULA (H_0)** ⇨ Es la afirmación contraria a la que ha llegado el investigador. Es la hipótesis que el investigador pretende rechazar. Si tiene la evidencia suficiente para ello, podrá probar que lo contrario es cierto.
- **HIPÓTESIS ALTERNATIVA (H_a)** ⇨ Es lo que el investigador realmente piensa que es la causa del fenómeno. Se acepta si se rechaza H_0. Es implícita.

Si hay evidencia para rechazar H_0 => se acepta H_a

Si no hay evidencia para rechazar H_0 => se "acepta" temporalmente H_0.

Scanned with CamScanner

* NIVEL DE SIGNIFICANCIA

El nivel de significancia α es la probabilidad que se deja fuera del intervalo de confianza de una distribución, permitiendo determinar con que probabilidad el estadístico de prueba se encuentra en la zona de rechazo o no.

Es decir que α es el complemento al intervalo de confianza.

Hay 2 tipos de errores:

CONCLUSIÓN / REALIDAD	Ho VERDADERA	Ho FALSA
Ho VERDADERA	✓	ERROR DE TIPO I
Ho FALSA	ERROR DE TIPO II	✓

- ERROR DE TIPO I ⇒ Ho es verdadera y el investigador la rechaza.

  La probabilidad de que ocurra un error de tipo I es:

  $P(\text{Rechazar Ho} \mid \text{Ho es V}) = \alpha$ → Esto es que el estimador $\hat{\sigma}$ caiga en la zona de rechazo erróneamente.

- ERROR DE TIPO II ⇒ Ho es falsa y el investigador la acepta.

  La probabilidad de que ocurra un error de tipo II es:

  $P(\text{Aceptar Ho} \mid \text{Ho es F}) = \beta$ → Esto es que el estimador $\hat{\sigma}$ caiga en la zona de aceptación erróneamente.

~~$\boxed{1-\beta = \text{potencia del test}}$~~

~~↳ Test deseable: $1-\beta > \alpha$~~

* ESTADÍSTICO DE PRUEBA

Un estadístico de prueba es una V.A. calculado a partir de los datos de la muestra y usado para la P.H. El estadístico de prueba compara los datos con lo que plantea la hipótesis nula.

* REGIÓN DE RECHAZO

Conociendo el estadístico de prueba y su distribución, se define en la distribución una región de rechazo de Ho.

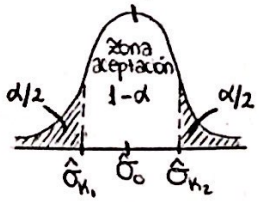
Esa región de rechazo es un intervalo en el que se cuenta con evidencia suficiente para rechazar la Ho con probabilidad α de cometer un error de tipo I.

Scanned with CamScanner

La forma de esta región de rechazo depende de como se halla planteado H_a:

$\bar{x}$ → es la distribución del estimador (ej: TLCMarce).

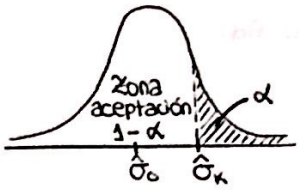
PRUEBA DE DOS COLAS



$H_a: \hat{\sigma} \neq \hat{\sigma}_0$

$RR = \{z \mid z_{\frac{\alpha}{2}} < z < z_{1-\frac{\alpha}{2}}\}$

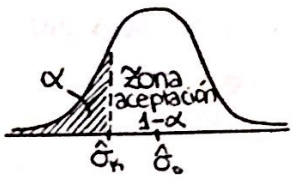
PRUEBA DE COLA DERECHA



$H_a: \hat{\sigma} > \hat{\sigma}_0$

$RR = \{z \mid z > z_{1-\alpha}\}$

PRUEBA DE COLA IZQUIERDA



$H_a: \hat{\sigma} < \hat{\sigma}_0$

$RR = \{z \mid z < z_{\alpha}\}$

Con $\hat{\sigma}_K$ el valor critico que separa la región de aceptación de la de rechazo. Se obtiene de tablas de probabilidad acumulada y depende de la distribución del estadistico y de α.

*VALOR P

Es un número entre 0 y 1 que indica la probabilidad de obtener en una P.H. un valor del estadistico de prueba posible, tomando como cierta la H_0.

Las PH se pueden hacer en función de los valores criticos o del valor p:

- Prueba de dos colas => Se rechaza H_0 si $\hat{\sigma}_{obs} < \hat{\sigma}_{K_1}$ o $\hat{\sigma}_{obs} > \hat{\sigma}_{K_2}$
  => " " " " valor-p $= 2 \cdot P(\hat{\sigma} > \hat{\sigma}_{obs} \mid H_0 \text{ es V}) < \alpha$
- Prueba de cola derecha => Se rechaza H_0 si $\hat{\sigma}_{obs} < \hat{\sigma}_K$
  => " " " " valor-p $= P(\hat{\sigma} > \hat{\sigma}_{obs} \mid H_0 \text{ es V}) < \alpha$
- Prueba de cola izquierda => Se rechaza H_0 si $\hat{\sigma}_{obs} > \hat{\sigma}_K$
  => " " " " valor-p $= P(\hat{\sigma} < \hat{\sigma}_{obs} \mid H_0 \text{ es V}) < \alpha$

PH PARAMÉTRICAS

Estas se usan cuando se ha seleccionado previamente un modelo para la MA. que obvio tiene parámetros que lo caracterizan. Entonces, se plantea una hipótesis relacionada con el valor de algún parámetro de la distribución teórica del estimador.

* PH PARA LA MEDIA

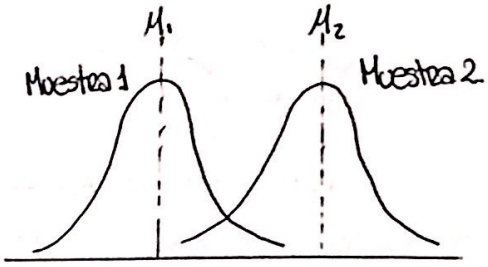
$H_0 \Rightarrow \mu = \mu_0$. Estimador: $\hat{\mu} = \bar{x}$

CONDICIONES	DISTRIBUCIÓN DE $\hat{\mu} = \bar{x}$	ESTADÍSTICO DE PRUEBA
• Distribución de la población: $N(\mu_0, \sigma^2)$ o no • Tamaño de la muestra: $n > 30$ • Desviación de la población: conocida	$\sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$	$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
• Dist. de la población: $N(\mu_0, \sigma^2)$ • Tamaño de la muestra: $n > 30$ • Desviación de la población: desconocida	$\sim N\left(\mu_0, \frac{S^2}{n}\right)$	$z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$
• Dist. de la población: $N(\mu_0, \sigma^2)$ • Tamaño de la muestra: $n < 30$ • Desviación de la población: desconocida.	$\sim t_{n-1}$	$t = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$

Scanned with CamScanner

Los pasos son:

1. Establecer H_0 y H_a.
2. Seleccionar el nivel de significancia.
3. Seleccionar el estadístico de prueba.
4. Definición de la RR.
5. Toma de desición.



*** PH PARA LA DIFERENCIA DE MEDIAS**

$H_0 \Rightarrow \mu_1 - \mu_2 = \Delta\mu$. Estimador. $\Delta\hat{\mu} = \Delta\bar{X} = \bar{X}_1 - \bar{X}_2$

CONDICIONES	DISTRIBUCIÓN DE $\Delta\bar{X}$	ESTADÍSTICO DE PRUEBA
Población $\sim N(\mu_i, \sigma_i^2)$ ↓ σ conocida o no y n grande	$\sim N\left(\Delta\mu, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$
Población $\sim N(\mu_i, \sigma_i^2)$ o no y n grande. σ desconocida	$\sim N\left(\Delta\mu, \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)$	$t =$ _____

*** PH PARA UNA PROPORCIÓN**

Bin: $\mu = n.p$, $\sigma^2 = n(1-p)p$, $\hat{p} = \frac{\bar{X}}{n}$

Para n grande $Bin(n,p) \rightsquigarrow N(np, npq)$

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1) \Rightarrow P\left(Z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

*** PH PARA LA VARIANZA**

El estadístico de prueba de σ^2 es S^2 con distribución χ^2. $H_0 \Rightarrow \sigma^2 = \sigma_0^2$

*** PH PARA LA COMPARACIÓN DE VARIANZAS**

Para comparar las varianzas de dos M.A., el estadístico de prueba es $\frac{S_1^2}{S_2^2}$ con distribución fea. $H_0 \Rightarrow \sigma_1^2 = \sigma_2^2$.

Scanned with CamScanner

PH NO PARAMÉTRICAS

No se conoce el modelo, por eso es no paramétrica. Por ejemplo, se tienen 2 MA. con PDF desconocidas y con histogramas parecidos, y se quiere saber si las dos muestras provienen de la misma población.

PRUEBAS DE BONDAD DE AJUSTE

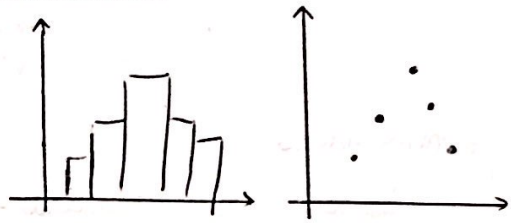
Indica una medida de bondad que expresa la discrepancia entre los valores observados y los valores esperados por una distribución teórica F.

H_0 = "los datos provienen de la distribución F".

H_a = "los datos no provienen de la dist. F".

*TEST CHI-CUADRADO

- Se aplica a distribuciones tanto continuas como discretas.
- Determina la independencia entre dos variables.
- Se calculan las "frecuencias observadas" y se las compara con las "frecuencias teóricas".



El estadístico con distribución χ^2:

$$\chi^2_{CHI} = \sum_i \frac{(f_{obs} - f_{teo})^2}{f_{teo}}$$

Se hace la prueba de hipótesis → H_0 = la dist. empírica sigue la dist. teórica.
→ H_a = la dist. empírica no sigue la dist. teórica.

Para una dada significancia $\alpha \Rightarrow 1-\alpha = Z$ y un nº de grados de libertad (si el histograma tiene n bines tiene n-1 gl). Se definen dos casos:

$\chi^2_{CHI} < \chi^2_{Z,gl} \Rightarrow$ se acepta H_0.

$\chi^2_{CHI} > \chi^2_{Z,gl} \Rightarrow$ se rechaza H_0.

$\chi^2_{Z,gl}$ es valor de tabla

También se puede calcular el valor-p con la CDF de χ^2_{CHI}:

$p = 1 - CDF_{\chi^2}$ → $p > \alpha \Rightarrow$ se acepta H_0.
→ $p < \alpha \Rightarrow$ se rechaza H_0.

*TEST DE KOLMOGOROV-SMIRNOV (KS)

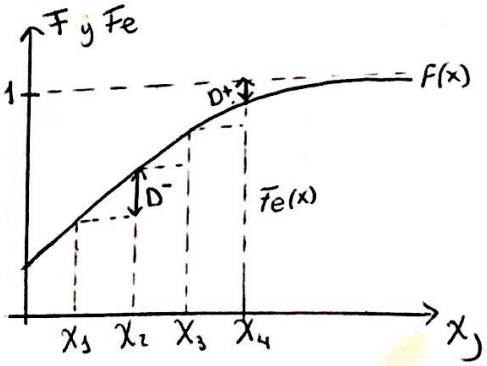
- Se aplica a distribuciones continuas
- Compara distribuciones acumuladas observadas (empíricas) y esperadas (teóricas).

Sea $F_e(x)$ la CDF empírica y $F(x)$ la CDF teórica, entonces H_0 es "$F_e(x)$ es cercana a $F(x)$".

El estadístico es $D = \max_x |F_e(x) - F(x)|$,

$$D^+ = \sup_x \{F_e(x) - F(x)\} = \max_{1 \le j \le n} \left\{\frac{j}{n} - F(x_j)\right\}$$

$$D^- = \sup_x \{F(x) - F_e\} = \max_{1 \le j \le n} \left\{F(x_j) - \frac{j-1}{n}\right\}$$



Scanned with CamScanner

Luego, $D = \max\{D^+, D^-\}$.

También, se puede calcular el valor-p:

$$\text{valor-p} = P_F(D \geq d) \rightarrow \begin{cases} < \alpha \Rightarrow \text{Rechaza } H_0. \\ > \alpha \Rightarrow \text{Acepta } H_0. \end{cases}$$

MAPA CONCEPTUAL 20

- INFERENCIA
  - BAYESIANA
  - FRECUENTISTA
    - Ejemplo: la probabilidad de éxito de un problema de Bernoulli se obtiene realizando el experimento un nº n de veces (moneda cara) y calculando el promedio de éxitos $\bar{x} = \frac{p}{n} = \frac{1+0+0+\dots+1+1}{n}$ y acá no entró Bayes.
  - NO PARAMÉTRICAS
    - ESTIMACIÓN DE LA PDF
      - HISTOGRAMAS
      - KDE
      - BLOQUES BAYESIANOS
      - MEZCLA DE GAUSSIANAS
  - PARAMÉTRICAS
    - EVIDENCIA Y PRIOR = CTES → MAXIMUM LIKELIHOOD → ERRORES GAUSSIANOS
      - CUADRADOS MÍNIMOS
        - MODELO LINEAL → La solución de los cuadrados mínimos es una operación cerrada.
        - MODELO NO LINEAL → Solución de cuadrados mínimos con métodos iterativos.
          - MÉTODO GRADIENTE DESC.
          - MÉTODO GAUSS-NEWTON.
          - MÉTODO LEVEM.-MARQ.
    - EVIDENCIA Y PRIOR ≠ CTES → FULL BAYES
      - MÉTODOS MONTECARLO (Maximizar full Bayes no suele tener solución analítica)
        - CADENAS DE MARKOV
          - ALGORITMO DE METRÓPOLIS-HASTINGS
          - ALGORITMO DE GIBBS.
  - SELECCIÓN DE MODELOS
    - ELEGIR EL MODELO
    - ELEGIR LA CANTIDAD DE PARÁMETROS
    - ENCONTRAR LOS MEJORES PARÁMETROS.
  - PREPROCESAMIENTO DE DATOS
    - TEST AND TRAIN SAMPLE → VALIDACIÓN CRUZADA
    - PROBLEMAS
    - REDUCCIÓN DE DIMENSIONALIDAD
      - TIRAR COMPONENTES
      - PCA
      - NON NEGATIVE FACTORIZATION
  - MACHINE LEARNING — TIPOS
    - NO SUPERVISADO → CLUSTERING
    - SUPERVISADO
      - REGRESIÓN
      - CLASIFICACIÓN
        - ÁRBOL DE DECISIÓN
          - BOOSTING
          - RANDOM FOREST
        - CLASIFICADOR NAIVE BAYES
        - EXPECTACIÓN/MAXIMIZACIÓN → MEZCLA DE GAUSSIANAS
        - MARGIN CLASSIFIERS
          - MAXIMUM MARGIN CLASSIFIERS
          - SOFT MARGIN CLASSIFIERS
          - SUPPORT VECTOR MACHINES
        - REGRESIÓN LOGÍSTICA.
        - MÉTRICAS
  - REDES NEURONALES

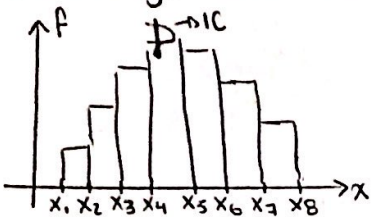
Scanned with CamScanner

INFERENCIA BAYESIANA

NO PARAMÉTRICAS → Existen métodos para determinar la función $\hat{f}(x)$ que estima la PDF.

HISTOGRAMAS

Un histograma es una función definida por partes, de manera que:



$$\hat{f}(x) = \begin{cases} C_1 & x_1 \leq x \leq x_2 \\ C_2 & x_2 < x \leq x_3 \\ \vdots & \vdots \\ C_m & x_m < x \leq x_{m+1} \end{cases}$$ con n intervalos

Esta función (el histograma) estima la PDF de una M.A.

Más formalmente, dada una $\{x_i\}_{i=1}^{n}$ M.A, un histograma es una función $h(x): \mathbb{R} \to \mathbb{N}$ tal que:

$$h(x) = \sum_{i=1}^{n} \theta(x) . C_i$$

con C_i=cte y la función escalón:

$$\theta(x) = \begin{cases} 1 & l < x < v \\ 0 & \text{c.c.} \end{cases}, \quad \overbrace{B = \bigcup_{k=1}^{m} B_k}^{\text{bines}} \text{ con } B_i = [l_i, v_i)$$

Una observación (una V.A) tiene dos posibilidades:

$$\left.\begin{array}{l} x_i \in B_k \text{ (éxito)} \\ x_i \notin B_k \text{ (fracaso)} \end{array}\right\} \text{se puede pensar como un experimento de Benoulli.}$$

donde la probabilidad de éxito p_k se puede estimar: $\hat{p}_k = \frac{C_k}{n}$ representando la fracción de elementos de la M.A. que caen en el bin k.

Entonces, se puede calcular un intervalo de confianza a partir de la distribución del estimador $\hat{p}_k$ que si bien es Binomial (n, p_k), si consideramos una muestra grande => Bin ~ Normal.

$$\hat{f}_i \sim N\left(\frac{C_i}{n}, \frac{S_i^2}{n}\right) = N\left(\frac{C_i}{n}, \frac{p_i q_i}{n}\right) = N\left(\frac{C_i}{n}, \frac{C_i}{n} \cdot \frac{(1 - C_i/n)}{n}\right)$$

Este intervalo de confianza funciona como una barra de error para cada bloque del histograma: $(C_i \pm \sigma_i)$ donde $\sigma_i = \sqrt{n_i}/\Delta.n$

La cantidad Δ es el ancho del bin. Si la distribución es gaussiana el ancho óptimo está dado por la regla de Scott:

$$\Delta = \frac{3.5\sigma}{N^{1/3}}$$

Pero si no es gaussiana, se usa la de Freedmann-Diaconis:

$$\Delta = \frac{2.7\sigma}{N^{1/3}}$$

Scanned with CamScanner

ESTIMACIÓN DE DENSIDAD POR KERNEL (KDE)

Se suele usar cuando se tienen pocos datos.

KDE es un método que permite estimar la PDF de una M.A. con una función resultante de la suma de muchas "funciones de Kernel".

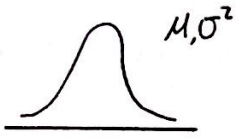
Consiste en representar cada observación x_i de una M.A. con una distribución llamada función de Kernel y luego, en lugar de contar cuántos elementos caen en un intervalo, se suman las funciones de Kernel, obteniendo una función resultante suave.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

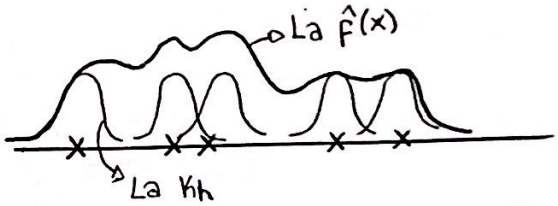
↳ función de Kernel que depende del parámetro de suavizado h (ej. Gaussiana).

La función de Kernel tiene previamente sus parámetros fijados, es decir:

K_h:



La muestra y su PDF:



NOTA Hay métodos para determinar el h óptimo → si es muy chico serían picos indistinguibles, si es muy grande se pierden detalles.

BLOQUES BAYESIANOS (BAYESIAN BLOCKS)

→ "Es un histograma irregular".

El método de bloques bayesianos se trata de construir una PDF para una M.A. a partir del conjunto de funciones determinadas en intervalos de ancho variable.

Los datos se segmentan en bloques, y los límites de los bloques se establecen mediante puntos de cambio. Usando un análisis bayesiano, dentro de cada bloque, se define la función llamada log-likelihood fitness function:

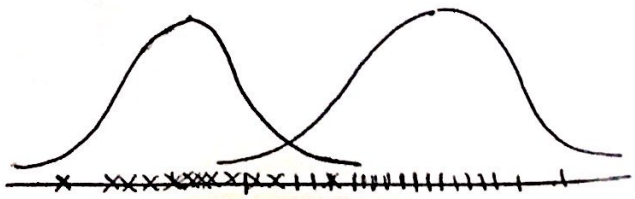
→ Con Bayes se determina la prob. de cada bloque

$$(N_i, T_i) = N_i\left[\mathrm{Log}(N_i) - \mathrm{Log}(T_i)\right]$$

donde N_i es el nº de puntos en el bloque i y T_i es el ancho del bloque i. Debido a la naturaleza aditiva de las probabilidades logarítmicas, la $\hat{f}_T$ es la suma de las $\hat{f}$ de cada bloque individual.

MEZCLA DE GAUSSIANAS (GAUSSIAN MIXTURE)

La mezcla de gaussianas es un método para estimar la $\hat{f}$ dada por la suma de gaussianas. Las gaussianas involucradas no deben tener los mismos parámetros necesariamente.



Scanned with CamScanner

PARAMÉTRICAS → se estudian los parámetros una vez definido el modelo de una M.A.

EVIDENCIA Y PRIOR CONSTANTES

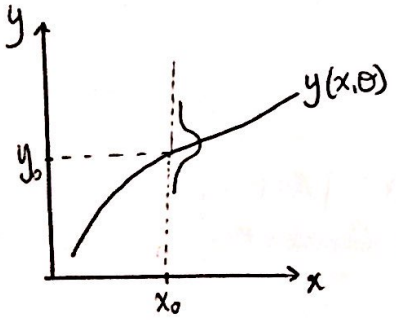
En inferencia Bayesiana, cuando la evidencia y el prior son ctes, la selección de parámetros que mejor describen los datos se hace con el método de MAXIMUM LIKELIHOOD (fundamento en hoja 8), donde los mejores estimadores surgen de maximizar el $\log(\mathcal{L})$.

En la mayoría de las muestras, los errores son **ERRORES GAUSSIANOS** y esto se debe a que, en general, las fuentes de error son muchas, haciendo el error total una "suma" de las distintas fuentes de error. Por el TLC, cuando una variable es suma de V.A. y se tiene una muestra grande, dicha variable tiene una distribución normal.

Si se quiere maximizar el $\log(\mathcal{L})$ y la muestra tiene errores gaussianos se usa ⇒ **CUADRADOS MÍNIMOS**

CUADRADOS MÍNIMOS

Sean $\{(x_i, y_i)\}_{i=1}^{n}$ un conjunto de observaciones con errores gaussianos. Queremos describir la correlación entre x e y con una relación $f(x,\theta)$, encontrando los parámetros que mejor representen la relación x, y.



Se puede pensar la f como normal centrada en $y(x,\theta)$ y con un valor de dispersión dado por la gaussiana que ajusta los errores:

$$f(x) \sim N(y(x,\theta),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[\frac{-(\overbrace{y_i}^{\text{dato}} - \overbrace{y(x_i;\theta)}^{\text{modelo}})^2}{2\sigma^2}\right] \quad (*)$$

Luego, el maximum likelihood: $\log(\mathcal{L}) = \sum_{i=1}^{n} \log\left(N(y_i, \sigma_i)(x_i)\right)$

Reemplazando la exponencial (*) dentro del logaritmo:

$$\log(\mathcal{L}) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum_{i=1}^{n} \frac{(y_i - y(x_i;\theta))^2}{2\sigma^2}$$

Usando $\log(\sqrt{2\pi\sigma}^{-1}) = -\log(\sqrt{2\pi\sigma})$:

$$\log(\mathcal{L}) = -n \cdot \log(\sqrt{2\pi\sigma}) - \sum_{i=1}^{n} \frac{(y_i - y(x_i,\theta))^2}{2\sigma^2}$$

Scanned with CamScanner

Si pensamos que maximizar el $\log(\mathcal{L})$ es equivalente a minimizar el $-\log(\mathcal{L})$:

$$-\log(\mathcal{L}) = \underbrace{n\log\left(\sqrt{2\pi\sigma}\right)}_{=cte} + \underbrace{\sum_{i=1}^{n} \frac{\left[y_i - y(x_i,\theta)\right]^2}{2\sigma^2}}_{sumatoria}$$

y minimizar [cte+sumatoria] es equivalente a encontrar el mínimo de la sumatoria.

Por lo tanto, buscamos los parámetros que minimizan los cuadrados de las diferencias entre los valores observados y los predichos:

$$\boxed{\theta_{optimos} = \min\left\{\sum_{i=1}^{n} \underbrace{\left[y_i - y(x_i,\theta)\right]}_{residuos}^2\right\}}$$

El método de cuadrados mínimos es un caso particular del método de maximum likelihood cuando los errores son gaussianos.

A la suma de residuos se le suele llamar función de costo o χ^2

<u>RESOLUCIÓN DE CUADRADOS MÍNIMOS</u>

Cómo resolver cuadrados mínimos depende de si el modelo es lineal o no.

*<u>MODELO LINEAL</u>

Si $y(x_i,\theta)$ depende de manera lineal de los parámetros, donde la f no es función de los parámetros: $y(x_i,\theta) = \sum_{k=1} a_k . f(x)$

Entonces, la función de costo:

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{y_i}{\sigma_i} - \frac{1}{\sigma_i}\underbrace{\sum_{k=1}^{M} a_k . f_k(x_i)}_{\text{lo que busco}}\right)^2$$

y se minimiza con <u>MÉTODOS MATRICIALES</u> que se escribe: $\chi^2 = |A - Bx|^2$ con $A = \frac{y_i}{\sigma_i}$ y $B = \frac{f(x_i)}{\sigma_i}$ matrices ctes y $x = \begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix} = a$.

A: dimensión M
B: dimensión $M \times n$

El mínimo va a estar en $\frac{\partial \chi^2}{\partial a_i} = 0 \ \forall a_i$, produciendo un sistema de M ec. una por cada parámetro:

$$0 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[y_i - \sum_{k=1}^{M} a_k f_k(x_i)\right] f_k(x_i) \quad k=1,\dots,M$$

y haciendo cuentas esto se puede reescribir como:

$$\sum_{j=1}^{M} \alpha_{kj} . a_j = \beta_k$$

Scanned with CamScanner

* MODELO NO LINEAL

Si $y(x,\theta)$ depende de manera no lineal de los parámetros no se puede aplicar el algebra matricial analítico y se usan <u>métodos iterativos</u>:

- Método del gradiente descendiente.
- Método de Gauss-Newton.
- Método de Levemberg-Marquardt.

} algoritmos de optimización.

○ MÉTODO DEL GRADIENTE DESCENDIENTE

Buscamos los parámetros que minimizan la función de costo, χ^2, cuando el modelo es no lineal, pero si se puede calcular la derivada.

El proceso de búsqueda es iterativo, se dan valores iniciales y se itera hasta que la derivada de χ^2 varíe poco:

$$\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \cdot \vec{\nabla}\chi^2(\vec{\theta}_i)$$

i: i-iteración

Me acerco al mínimo "dando saltos" proporcionales al gradiente de χ^2 y al η (tasa de aprendizaje). Saltos que son opuestos a la dirección de máximo crecimiento (por esto es que sé que voy hacia el mínimo).

Cerca del mínimo el gradiente varía poco (~ 0) y allí puedo frenar la iteración.

Es muy sensible al valor de η:

- η grande $\Rightarrow$ me acerco rápido al mínimo, pero al estar cerca oscila y se dificulta la convergencia.
- η chico $\Rightarrow$ se avanza poco y son necesarios muchos pasos, por lo que la convergencia puede ser muy lenta si estoy lejos.

○ MÉTODO DE GAUSS-NEWTON

Si escribimos la Serie de Taylor de χ^2 alrededor de $\vec{\theta}_k$:

varios parámetros $\left\{ \chi^2(\vec{\theta}) \simeq \chi^2(\vec{\theta}_k) + \left(\vec{\nabla}\chi^2(\vec{\theta}_k)\right)^T \cdot (\vec{\theta}-\vec{\theta}_k) + \frac{1}{2}(\vec{\theta}-\vec{\theta}_k)^T H(\vec{\theta}_k) \cdot (\vec{\theta}-\vec{\theta}_k) \right.$

parámetro único $\left\{ \chi^2(\theta) \simeq \chi^2(\theta_k) + \frac{d\chi^2}{d\theta}(\theta-\theta_k + \frac{1}{2}\frac{d^2\chi^2}{d\theta^2}(\theta-\theta_k)^2 \right.$

Una buena aproximación del mínimo es tal que la derivada primera (el gradiente) se anule y la derivada segunda (la hessiana) sea positiva. Luego, para implementarlo, es necesario calcular la hessiana.

- Hace saltos que dependen de la derivada segunda.
- Funciona bien cerca del mínimo.
- Lejos del mínimo la aproximación no es buena y puede generar saltos que lo alejen del mínimo.
- Problema: puede caer en mínimos secundarios.

Scanned with CamScanner

MÉTODO DE LEVEMBERG-MARQUARDT

→ El λ (parámetro de regularización) indica la suavidad de cambio (en caso de cambiar) o el peso de cada método (si considero una suma de ambos en todo momento).

Combina los dos métodos anteriores:

- Gradiente descendiente ⇒ lo usa lejos del mínimo para llegar rápido a la zona cercana al mínimo con η grande.
- Gauss-Newton ⇒ lo usa cerca del mínimo ya que este método se comporta bien allí.

EVIDENCIA Y PRIOR NO CONSTANTES

Cuando la evidencia y el prior no son ctes, encontrar la probabilidad posterior con una solución analítica en el teorema de Bayes es muy complicado.

↓ "maximizar $P(M|D)$"

$$\underbrace{P(M|D)}_{\text{posterior}} = \frac{\overset{\mathcal{L}}{P(D|M)} \cdot \overset{\text{prior}}{P(M)}}{\underbrace{P(D)}_{\text{evidencia}}} \Bigg\} \text{ Full Bayes}$$

Por lo que $\boxed{\text{se recurre a métodos numéricos}}$

El método numérico más utilizado es el MÉTODO MONTECARLO, usado para aproximar expresiones matemáticas complejas y costosas de evaluar con exactitud. Ver pág 6.

Los métodos MC proporcionan soluciones aproximadas con la realización de experimentos con muestreos de números pseudoaleatorios en una computadora.

A diferencia de los métodos numéricos que se basan en evaluaciones en N puntos en un espacio M-dimensional para producir una solución aproximada, MC tiene un error absoluto de la estimación que decrece como $\frac{1}{\sqrt{n}}$ (TLC).

> Lo que se busca es una medida de la tendencia central de la distribución de un parámetro, es decir, se quiere estimar el valor de expectación. El estimador de la media es el PROMEDIO.

ESTIMACIÓN DE VALORES DE EXPECTACIÓN

Como planteo general, queremos evaluar el valor de expectación de una función $f(z)$ de una V.A. que tiene distribución $p(z)$:

$$E[f] = \int f(z)\, dp(z)$$

Si tengo un conjunto de M.A. tomadas de la distribución $p(z)$, entonces

de L M.A.

$$\hat{f} = \frac{1}{L} \sum_{i=1}^{L} f(z_i) \Bigg\} \text{ estimador Montecarlo de } E[f]$$

Scanned with CamScanner

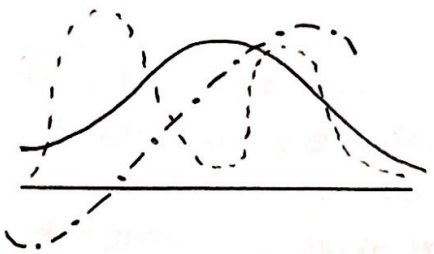
y se pueden usar varios metodos de muestreo para hacer esto:

- Método Importance sampling.
- MCMC
- etc.

* IMPORTANCE SAMPLING

Este método es "similar" al método de aceptación-rechazo, pero se usa cuando la distribución es más complicada.

Importance sampling permite estimar el valor de expectación, sin mediar el sorteo de la distribución $p(z)$, proponiendo una (PROPOSAL DISTRIBUTION) $q(z)$ de la cual sí es fácil sortear randoms.



$$E[f] = \int f(z)\, p(z)\, dz$$
$$= \int f(z) \cdot \frac{p(z)}{q(z)}\, q(z)\, dz$$
$$= \frac{1}{L} \sum_{i=1}^{L} \underbrace{\frac{p(z_i)}{q(z_i)}}\, q(z_i)$$

z_i: i-ésima muestra.

son los pesos que compensan el hecho de que estamos sorteando randoms de la dist. equivocada.

Para que este método funcione, debe ser fácil de evaluar $p(z)$ en los valores de muestreo $\{z\}_{i=1}^{L}$ y la $q(z)$ debe ser parecida a $p(z)$.

* CADENAS DE MARKOV CON MÉTODOS MONTE CARLO (MCMC)

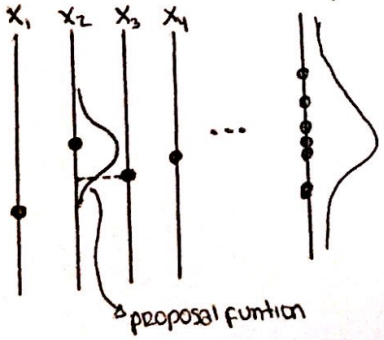
Este método permite obtener una secuencia de puntos que siguen la distribución de la probabilidad posterior. Dicha secuencia no es independiente, sino que cada punto depende del anterior, formando una especie de cadena.

Se usan MCMC para estimar el valor de expectación de cada parámetro:

$$E[\theta] = \int \theta\, P(\theta|D)\, d\theta \simeq \frac{1}{M} \sum_{i=1}^{M} \theta^{(i)}$$

con $\theta^{(i)}$ muestras de la cadena que muestrean el espacio de parámetros.

Acá también se usa una *proposal funtion*, pero no se sortea cada vez un nº random nuevo de la *proposal*, sino que se genera una a partir del anterior.



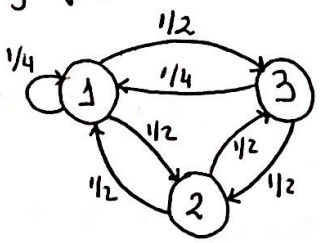
- De la proposal funtion, sorteé y conseguí el siguiente (de x_2 a x_3)
- Al final, aplasté todos los números (marginalización) y de ahí obtuve la distribución de esa muestra de no randoms.

Scanned with CamScanner

• **ESPACIO DE ESTADOS DISCRETOS** → se recorre el grafo y despues de mucho recorrer voy a visitar cada nodo con una frec prop al peso.

Una cadena de Markov es un conjunto $\{X_n, n \geq 0\}$ para el cual existen P_{ij} que representan la probabilidad de pasar de un estado a otro. P es la matriz de transición. Veamos ejemplo:

Hay 3 estados posibles para la V.A. X y se puede ir de uno a otro con ciertas probabilidades



$$P = \begin{array}{c} \\ \\ \\ \end{array} \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix} \begin{array}{l} \}=1 \\ \}=1 \\ \}=1 \end{array}$$

- Una cadena de Markov es irreducible si para cada par de estados existe una probabilidad no nula de ir de uno al otro.
- Una cadena de Markov es aperiódica si

Si la cadena es irreducible y se recorre el espacio por mucho tiempo, se puede demostrar que existen valores constantes π_j que indican la probabilidad de que el sistema esté en el estado j: son las PROBABILIDADES ESTACIONARIAS. Para las probabilidades estacionarias se cumple que al aplicar la matriz de transición se obtiene la misma distribución:

$$\pi = \pi P$$

es decir, la probabilidad de que el sistema esté en cierto estado no cambia a pesar de seguir "caminando" la cadena. En ese caso, se dice que la cadena está quemada.

peso nodo ≡ valor prob posterior

• **ESPACIO DE ESTADOS CONTINUOS**

Para construir cadenas de Markov en espacios de estados continuos hay distintos algoritmos:

- Algoritmo de Metrópolis-Hastings.
- Algoritmo de Gibbs.

① ALGORITMO DE METRÓPOLIS-HASTINGS

Combina ideas: *MonteCarlo* (saltos randoms con ciertas probabilidades), Markov (cada valor depende del anterior) y *Importance sampling* (sorteo de una función más fácil)

Scanned with CamScanner

El algoritmo consiste en elegir un punto arbitrario x_t que será el primer valor de la muestra y elegir una densidad de probabilidad arbitraria, $g(x|y)$, una proposal function, que sugerirá un candidato para el siguiente valor de la muestra, x, dado el valor anterior, y.

Una opción habitual es dejar que g sea una distribución gaussiana centrada en y, así los puntos cercanos a y serán más propensos a ser visitados a continuación, convirtiendo la secuencia de muestras en una caminata aleatoria.

Cuando se propone un x' de la distribución $g(x'|x_t)$, se define:

$$\alpha = \frac{P(x')}{P(x_t)}$$ ¿Cómo se calcula esa prob?

y se acepta o se rechaza x' a partir de:

- Generar un número aleatorio uniforme $\mu \in [0,1)$.
- Si $\mu \leq \alpha$, entonces se acepta el candidato de forma que $x_{t+1} = x'$.
- Si $\mu > \alpha$, entonces se rechaza el candidato y coloca a $x_{t+1} = x_t$ en su lugar

Así, este algoritmo continúa intentando moverse aleatoriamente por el espacio muestral, a veces aceptando movimientos y a veces no.

② ALGORITMO DE GIBBS

Con este algoritmo, en lugar de dar un salto en el espacio de parámetros, se marginaliza en todos los parámetros menos 1 y salta solo en ese parámetro.

Después de varios saltos, se tiene una distribución marginal para cada parámetro.

Se pueden recorrer los parámetros en orden o de forma aleatoria

---

MCMC NO es un método: de optimización
de maximización
de ajuste

es algoritmo de muestreo

MCMC es un método de estimación del valor de expectación para aproximar la probabilidad posterior.

Scanned with CamScanner

SELECCIÓN DE MODELOS

La idea es encontrar el modelo (mejor ajuste) que resulta más adecuado para representar los datos. Para encontrar dicho modelo hay que:

1. Elegir una familia de modelos en base a alguna teoría o seleccionar un modelo a partir de la exploración descriptiva de los datos.
2. Obtener los parámetros que mejor ajustan las observaciones.
3. Hacer una prueba de hipótesis para ver si los parámetros seleccionados están en buen acuerdo con los datos con algún nivel de significancia.
4. Repetir todo con varias familias de modelos para ver si hay otro modelo que ajuste mejor.

Un buen criterio para elegir el modelo es seleccionar aquel que logra un balance entre la varianza y el sesgo.

SESGO

El sesgo o bias se produce porque el modelo asume una serie de simplificaciones.

- alto sesgo => supone demasiadas simplificaciones para el modelo.
- bajo sesgo => no supone muchas simplificaciones y el modelo es más complejo.

El sesgo tiene que ver con cuanto se corre el modelo respecto a los datos.

VARIANZA

La varianza indica cómo cambia la función de costo (residuos, diferencia entre el modelo y los datos) cuando se utiliza un conjunto de datos en lugar de otro.

Si cambio el conjunto de datos y quiero ver que tan bien reproduce el modelo al nuevo conjunto de datos. Si la función de costo cambia poco al cambiar los datos, entonces el modelo tiene baja varianza.

Si el modelo ajusta las fluctuaciones particulares de un conjunto de datos, el modelo tendrá alta varianza.

COMPROMISO SESGO-VARIANZA

↑ SESGO ↓ VARIANZA
↓ SESGO ↑ VARIANZA

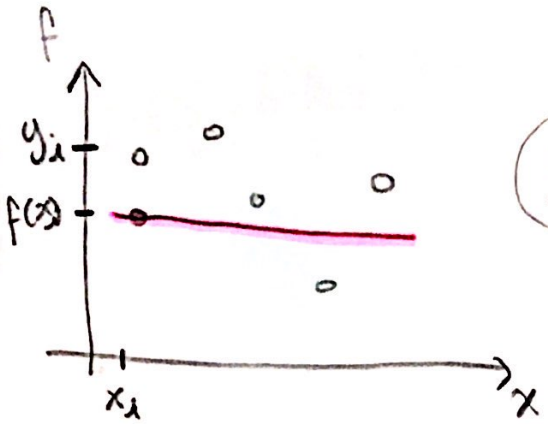
En general, cuando disminuye uno aumenta el otro.

Esto se debe a que a medida que se aumenta la complejidad del modelo (es decir, disminuyo el sesgo), el modelo ajusta mejor las fluctuaciones particulares del conjunto de datos usado (aumenta la varianza).

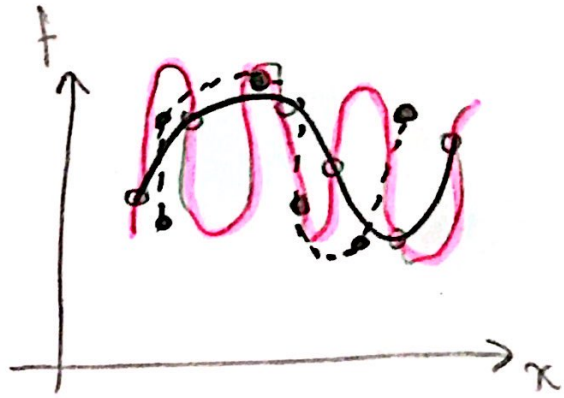
Para medir que tanto un modelo se parece a un conjunto de datos, se define el ERROR CUADRÁTICO MEDIO:

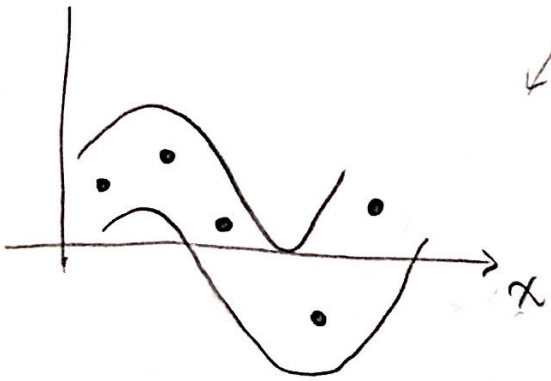
$$MSE = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(y_i - M(x_i; \vec{\theta})\right)}_{\text{residuos}}^2$$

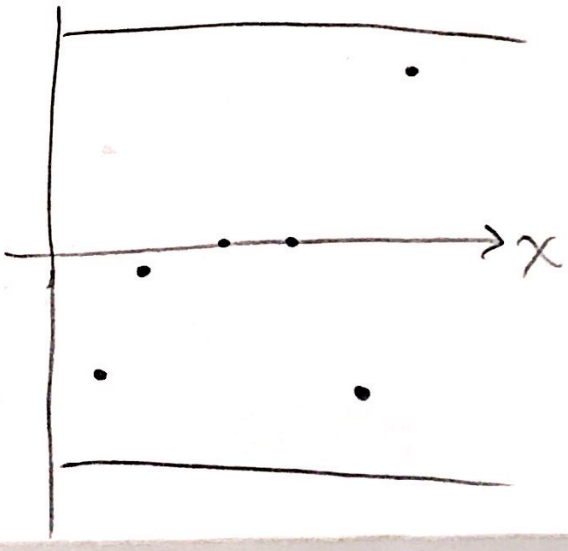
Scanned with CamScanner



$y_i - f(x_i)$







puntos no curvas

↳ predicciones del modelo.

Metodos no random.
Max. Lik. y cuada.
Compn. sesgo varianza
redes neuronales.

Scanned with CamScanner

SELECCIÓN DE MODELOS

La idea es encontrar el modelo (mejor ajuste) que resulta más adecuado para representar los datos. Para encontrar dicho modelo hay que:

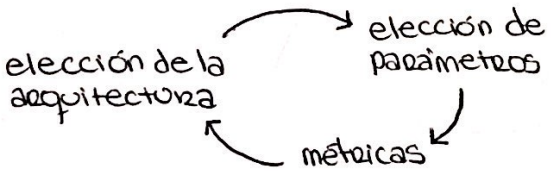
1. Elegir una familia de modelos en base a alguna teoría o seleccionar un modelo a partir de la exploración descriptiva de los datos.
2. Obtener los parámetros que mejor ajustan las observaciones.
3. Hacer una prueba de hipótesis para ver si los parámetros seleccionados están en buen acuerdo con los datos con algún nivel de significancia.
4. Repetir todo con varias familias de modelos para ver si hay otro modelo que ajuste mejor.

Penalización del Likelihood => penaliza la complejidad del modelo

↳ ① penalizar la cantidad de parámetros.

↳ ② penalizar el valor que toman los parámetros. } se usan diferentes pesos.

* autoML => elije automáticamente los mejores parámetros.
* MLOPS => el usuario repite varias veces:



Esto se debe a que a medida que se aumenta la complejidad ... el sesgo), el modelo ajusta mejor las fluctuaciones particulares del conjunto de datos usado (aumenta la varianza).

Para medir que tanto un modelo se parece a un conjunto de datos, se define el <u>ERROR CUADRÁTICO MEDIO</u>:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(y_i - M(x_i;\vec{\theta})\right)}_{\text{residuos}}^2$$

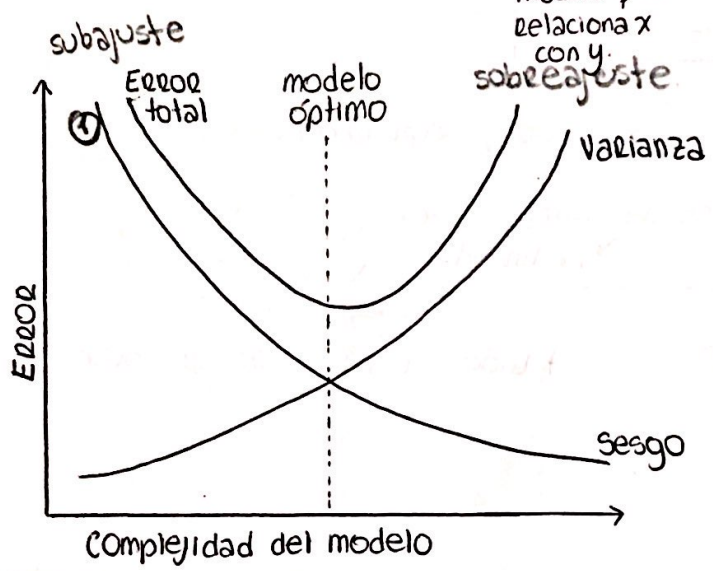
Scanned with CamScanner

donde el valor de expectación del MSE se puede escribir como la suma de 3 cantidades fundamentales:

$$\boxed{E\left[\left(y_i - M(x_i;\vec{\theta})\right)^2\right] = \underbrace{Var\left(M(x_i;\vec{\theta})\right)}_{\text{varianza del modelo}} + \underbrace{bias\left(M(x_i;\vec{\theta})\right)^2}_{\text{sesgo del modelo}} + \underbrace{Var(\varepsilon)}_{\text{varianza del error}}}$$

compromiso entre sesgo y varianza.

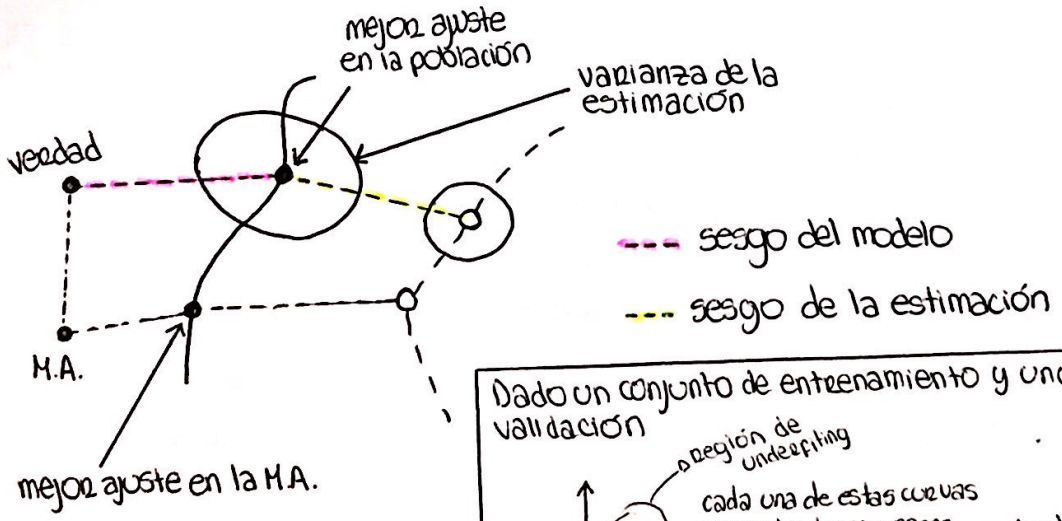
donde asumí que el modelo es $y = \underbrace{M(x_i,\vec{\theta})}_{\text{modelo que relaciona } x \text{ con } y} + \underbrace{\varepsilon}_{\text{error.}}$



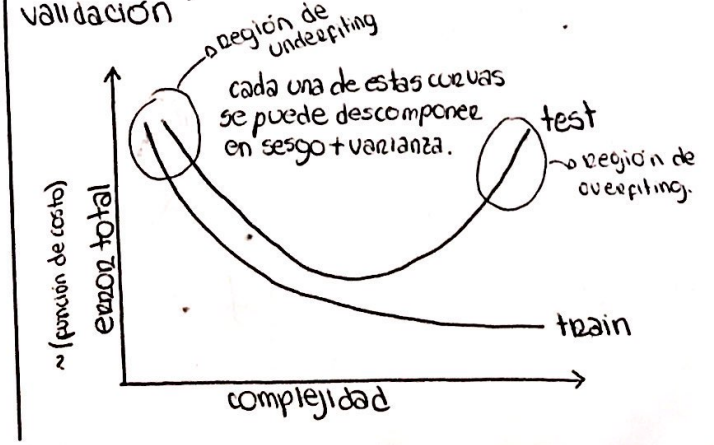
curva de aprendizaje.

me quedo con el modelo donde la curva de error total alcanza el mínimo.

ESPACIO DE LOS MODELOS



Dado un conjunto de entrenamiento y uno de validación



1

MODELOS ANIDADOS

Se dice que dos modelos son anidados cuando el modelo más simple es un caso particular del modelo más complejo.

Ejemplo:

- Un polinomio de orden 5 y uno de orden 6 son modelos anidados. Allí, la complejidad del modelo se mide con la cantidad de parámetros.
- Un polinomio y un seno no son modelos anidados. No es tan sencillo comparar sus complejidades.

CRITERIOS PARA DETERMINAR EL MEJOR MODELO

ANIDADOS → armar la curva de error y elegir la complejidad

NO ANIDADOS → criterios de → información. → Akaike, Fisher, Bayesiana } preguntar qué saben de esto.

→ likelihood.

Se basa en "penalizar" el likelihood:
si M_1 está anidado dentro...

+ Ridge y LASSO Regression.

Hablar esto con Marce. (NOTEBOOK 15)

Scanned with CamScanner

PREPROCESAMIENTO DE DATOS

TEST & TRAIN SAMPLE → Se separa la muestra en dos partes:

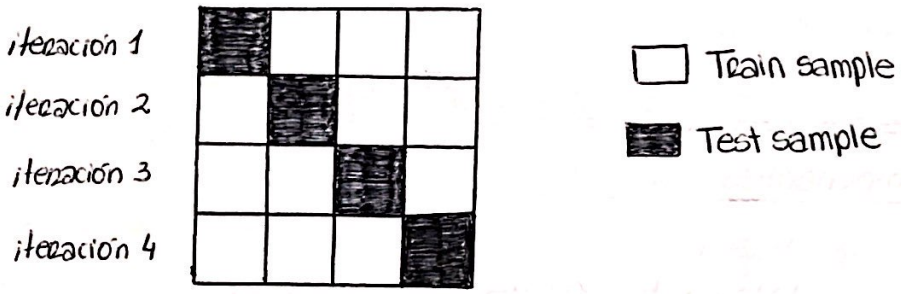
CONJUNTO DE ENTRENAMIENTO => Conjunto de datos con el que se le enseña a la computadora un modelo.

CONJUNTO DE VALIDACIÓN => Conjunto de datos con el que se testea el modelo aprendido por la computadora.

VALIDACIÓN CRUZADA

La división de los datos en dos conjuntos se puede hacer de muchas maneras, en particular, con validación cruzada:

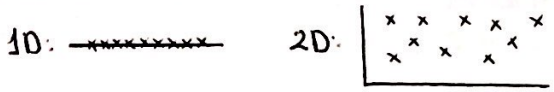
Se divide la muestra en K subconjuntos, uno se usa como test sample y el resto (K-1) como train sample. Esto se repite en K iteraciones (va entrenando y testeando varias veces el mismo modelo).



Este método de separación de los datos garantiza que los resultados obtenidos son independientes de la partición entre datos de entrenamiento y de validación.

PROBLEMAS

- "Maldición de la dimensionalidad": al aumentar el número de dimensiones, el volumen del espacio crece y la distribución de los datos se vuelve escasa (los puntos se esparcen).



- "No hay almuerzos gratis": no existe un modelo que funcione bien para todos los problemas, ya que no existe en la teoría una estrategia de optimización general. Cada situación debe determinar su mejor modelo.

Scanned with CamScanner

REDUCCIÓN DE DIMENSIONALIDAD

A raíz de la "maldición de la dimensionalidad" es necesario hacer reducción de dimensionalidad. Son técnicas de aprendizaje no supervisado.

Esta reducción puede hacerse de dos maneras:

1. Eliminando features (que son redundantes o no aportan información).
2. Transformando a un espacio de features de menor dimensión. Existen proyecciones dentro de los datos que capturan las principales correlaciones físicas y estadísticas entre las features.

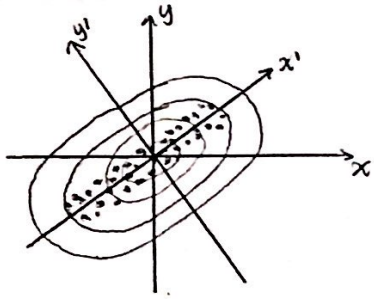
En el segundo caso, pueden ser transformaciones lineales o no lineales.

La reducción de dimensionalidad permite reducir el costo y el tiempo de proceso, pero se debe tener en cuenta que se tienen que incluir un cjto suficiente de características para lograr un alto rendimiento; en definitiva, encontrar el conjunto óptimo.

TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD LINEAL

- Análisis de componentes principales (PCA)

Es un algoritmo que busca la proyección ortogonal de los datos en un subespacio tal que la varianza de los datos proyectados es máxima.



Las componentes principales son las direcciones en las cuales los datos originales maximizan la dispersión. Es decir que los features (o combinación de ellos) más interesan-tes son los que tienen mayor variación.

Veamos un ejemplo; en el dibujo se ve una distribución bidimensional de puntos de una gaussiana centrada en el origen de los ejes x e y. La mayor correlación no se alinea con la elección inicial de los ejes x e y. En este caso x' es una componente principal.

Si deseamos reducir la cantidad de features (la cantidad de ejes) que se usan para describir estos datos, se deben rotar los ejes para alinearnos con esta correlación. Si la rotación maximiza la varianza a lo largo de los ejes resultantes se está haciendo análisis de componentes principales (PCA).

¿Cómo se obtiene la dirección de máxima varianza?

Dado un conjunto de datos, $\{x_i\}$, de N datos con N features se construye la matriz de covarianza:

Scanned with CamScanner

$$\Sigma = \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_n) \\ Cov(x_2,x_1) & Var(x_2) & \cdots & Cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_n,x_1) & Cov(x_n,x_2) & \cdots & Var(x_n) \end{bmatrix}$$

donde las componentes fuera de la diagonal distintos de cero dentro de la matriz de covarianza surgen porque existen correlaciones entre las características medidas. PCA identifica una proyección del conjunto de datos, $\{x_i\}$, que está alineado con las direcciones de máxima varianza. Se resuelve mediante un problema de autovalores y autovectores. → Se hace una transformación (rotaciones) para obtener que la Σ transformada sea lo más diagonal posible obteniendo:

- Autovalores: pesan la varianza de cada dirección.
- Autovectores: son las nuevas direcciones.

- Factorización de matrices no-negativas (NMF)

En PCA las componentes principales pueden ser positivas o negativas. Por el contrario, para muchos sistemas físicos sabemos a priori que los datos son positivos, por ejemplo, el espectro de una galaxia es una suma lineal de componentes estelares.

La factorización de matriz no-negativa aplica una restricción adicional a los componentes principales, haciendo que sean positivas.

- Análisis de componentes independientes (ICA)

Es una técnica computacional para resolver lo que se conoce como el "party problem": si hay varios micrófonos en una habitación que contiene N personas, cada micrófono capta una combinación lineal de las N voces. El objetivo de ICA es utilizar el concepto de independencia estadística para aislar (o desmezclar) las señales individuales (ej. espectro de galaxia como combinación de espectros estelares).

El principio de ICA es que las señales de entrada deben ser estadísticamente independientes. Dos variables aleatorias se consideran estadísticamente independientes si su distribución de probabilidad conjunta $f(x,y)$ puede describirse como la multiplicación de sus probabilidades individuales.

TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD NO LINEAL

- Aprendizaje múltiple

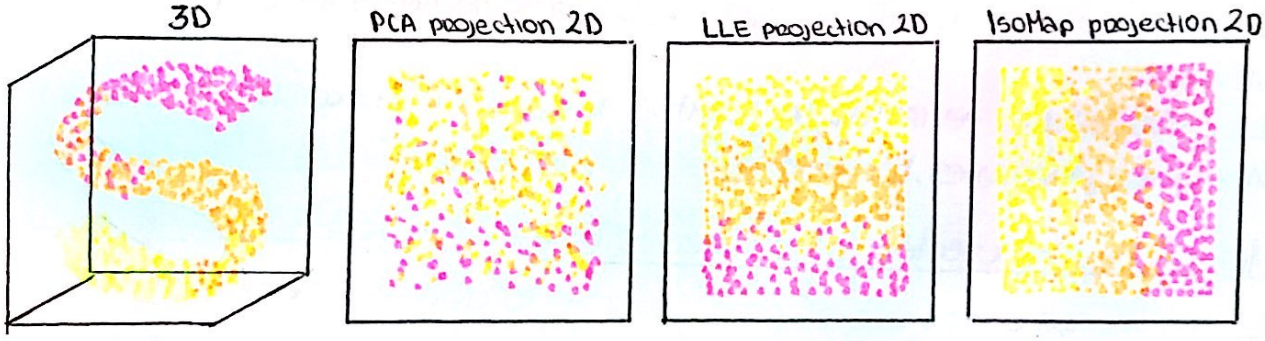
PCA, ICA y NMF son muy útiles, pero en general los conjuntos de datos pueden tener características no lineales. Ejemplo: las galaxias quiescentes pueden describirse bien mediante relativamente pocos componentes principales, pero las galaxias de núcleo activo poseen líneas de emisión y

Scanned with CamScanner

pueden requerir hasta ~30 componentes lineales para caracterizarse por completo. Estas líneas de emisión son características no lineales de los espectros y se requieren métodos no lineales para proyectar esa información en menos dimensiones.

El aprendizaje múltiple permite hacer reducción de dimensionalidad no lineal.

Supongamos un conjunto de datos que forma la figura de S 3D y queremos pasar a 2D; en este caso, PCA sale mal, ya que no hay proyección lineal en la que las partes distantes de la variedad no lineal no se superpongan. Sí se logra una buena proyección 2D con técnicas no lineales como la LLE o IsoMap.



* Incrustación lineal local (LLE)

Es un algoritmo que incrusta datos de alta dimensión en un espacio de menor dimensión conservando la geometría de barrios locales de cada punto. Estos vecindarios locales están determinados por la relación de cada punto con sus k vecinos más cercanos.

El primer paso es derivar un conjunto de pesos para cada dato que mejor reconstruyen el punto a partir de sus k vecinos más cercanos. Estos pesos codifican la geometría local de cada vecindario. En el segundo paso, con estos pesos mantenidos fijos, se encuentra un nuevo conjunto de datos de menor dimensión que mantiene las relaciones de vecindad descritas por estos pesos.

* Mapa isométrico (IsoMap)

Isomap recupera la estructura no lineal en menor dimensión aproximando las curvas geodésicas que se encuentran dentro de la variedad incrustada y calculando las distancias entre cada punto en el conjunto de datos a lo largo de estas curvas geodésicas. Necesito mejor explicación.

Scanned with CamScanner

Debilidades del aprendizaje múltiple

- No funciona bien con datos muy ruidosos o con datos que tienen "huecos".
- No hay técnicas para elegir el conjunto óptimo de k vecinos. La elección óptima en general depende de la densidad local de cada punto o de la curvatura de la variedad en cada punto.

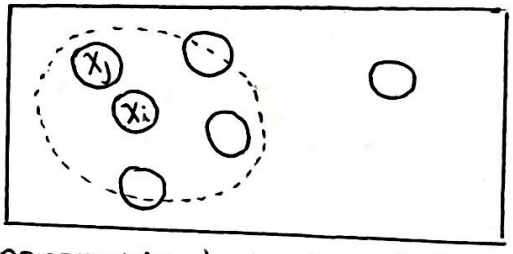
• t-distributed Stochastic Neighbor Embedding (t-SNE)

Es una técnica de reducción de dimensionalidad no lineal en la que se convierten distancias euclídeas en N dimensiones en probabilidades condicionales que representan "similaridad" (una medida de qué tan parecido es un dato a otro). Esta similaridad es la probabilidad de que un dato x_i elija a otro x_j como su vecino, sorteando todos los vecinos en base a una distribución t-Student de la distancia euclídea.

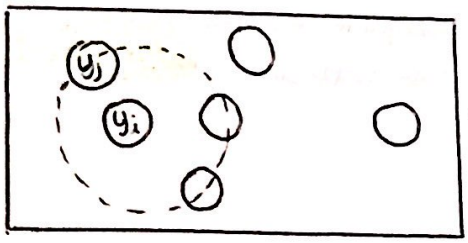
NOTA Existe la técnica Stochastic Neighbor Embedding (SNE) donde, en lugar de una distribución t-Student, la distribución es Gaussiana.

Los pasos de esta técnica son:

1. Encontrar la similitud por pares entre puntos cercanos en un espacio dimensional alto, convirtiendo las distancias euclideanas de alta dimensión entre los puntos de datos x_i y x_j en probabilidades condicionales $P(x_j|x_i)$.



x_i elige x_j como su vecino en función de la proporción de su densidad de probabilidad bajo una t-Student centrada en el punto x_i. Para los puntos de datos cercanos P es alta y para puntos muy separados P es baja.



2. Asignar cada punto en el espacio de alta dimensión a un mapa de baja dimensión basado en la similitud por pares de los puntos en el espacio de alta dimensión.

y_i e y_j son las contrapartes de baja dimensión de los puntos de datos de alta dimensión x_i y x_j. Se calcula la probabilidad condicional $Q(y_j|y_i)$ centrada bajo una t-Student en el punto y_i.

3. Encontrar una representación de datos de baja dimensión que minimice la falta de coincidencia entre P y Q usando un gradiente descendiente.

Scanned with CamScanner

MACHINE LEARNING

Se crean algoritmos donde se proponen arquitecturas y, a partir del aprendizaje de un conjunto de datos, se encuentra el mejor modelo que describe los datos.

El objetivo de los algoritmos de ML es encontrar una función para una variable de salida (y) dado un conjunto de variables de entrada (x). Esta función es la función de costo (f), la cual no es perfecta y produce un error en un conjunto de datos nuevos llamado error de predicción o de validación.

TIPOS DE ML

Existen dos tipos de variables:

- Variables objetivo => Es la variable (resultado) que se quiere predecir.
- Variables predictivas/explicativas => Son variables que caracterizan a los datos y permiten estimar la variable objetivo.

* Si se conoce cuál es la variable objetivo se hace ML SUPERVISADO, donde si la variable objetivo es continua (es decir que el resultado a predecir es un atributo numérico, ej: redshift) se usan técnicas de regresión; y si la variable objetivo es categórica (es decir que separa los datos en clases) se usan técnicas de clasificación.

* Si no se conoce cual es la variable objetivo se hace ML NO SUPERVISADO. Como no se sabe qué es lo que se quiere predecir, las técnicas de ML no supervisado describen las estructuras de los datos para encontrar grupos. usando técnicas de clustering.

ML NO SUPERVISADO

CLUSTERING

Cuando no se conoce el número, ubicación, tamaño o morfología de los grupos se pueden usar técnicas de agrupamiento para buscar grupos en el conjunto de datos basados en la proximidad de los objetos determinada por alguna medida de distancia o similaridad con un modelo.

- K-MEANS

Uno de los algoritmos más simples para clustering es k-means. Este agrupa datos semejantes en un número predefinido de grupos o clusters.

Cada grupo se representa por un "centroide", que es el promedio pesado de los elementos que contiene.

Scanned with CamScanner

Entrada: datos en un espacio n-dimensional.
Salida: solución con k grupos de datos.
Objetivo: minimizar la distancia entre datos/objetos de un mismo grupo.

El algoritmo k-means puede ser hard (cada dato pertenece a un clúster) o soft (cada dato tiene una probabilidad de pertenecer a uno u otro grupo).

Algoritmo: Se toman k centroides iniciales (en gral, aleatorios) y se le asigna a cada centroide el objeto más cercano, creando un clúster.
Luego, se encuentra el centroide de los objetos de un clúster y se asigna nuevamente a cada centroide el objeto más cercano.
Esto se realiza de forma iterativa y, a medida que se realizan más iteraciones, las agrupaciones se vuelven más estables y los centroides no se mueven. El algoritmo se repite hasta que se alcanza un criterio de detención (convergencia o nº máximo de iteraciones).

Medida de similaridad: la distancia define si dos elementos están "cerca" o no.

- ALGORITMO DE GONZÁLEZ

Este algoritmo se basa en minimizar el radio máximo de un grupo.
Se comienza sin grupos y se agrega progresivamente un grupo a la vez. Se selecciona arbitrariamente un centroide y se busca el punto que maximiza la distancia desde los centros de los grupos existentes y toma ese punto como el centroide del próximo grupo. Este procedimiento se repite hasta lograr k grupos. Luego, a cada objeto se le asigna la etiqueta de su centroide más cercano.

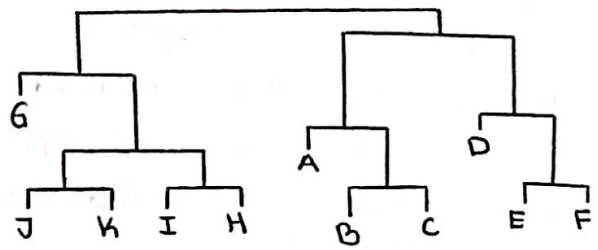
- MEAN SHIFT

Este algoritmo define grupos asociando cada dato a su pico más cercano. La idea de mean shift es mover los datos en la dirección del gradiente de densidad de los datos, esto lo hace encontrando los picos en una estimación de densidad por kernel (KDE). El número de grupos, k, es encontrado implícitamente por el método.

Scanned with CamScanner

• CLUSTERING JERÁRQUICO → A diferencia del arbol de desición, esto es no supervi-
-sado y cadena jerarquías solo. (galaxias, cúmulos, supcúm)

Este algoritmo relaja la necesidad de especificar el número de grupos. Para empezar, se dividen los datos en N grupos, uno para cada punto del conjunto de datos. Luego, se unen dos de los grupos dando como resultado N-1 grupos. Este procedimiento se repite hasta que la partición N contenga un clúster. Si dos puntos están juntos en el nivel M y permanecen juntos en todos los niveles posteriores, se conoce como agrupación jerárquica.



ML SUPERVISADO

REGRESIÓN → Toda técnica de regresión se puede categorizar (rangos) y hacer de clasificación.

Se desea inferir el valor esperado de y dado x (el valor condicional). Si se tiene un modelo para la distribución condicional (descrito por los parámetros θ) se puede escribir esta función como $y = f(x|\theta)$, donde x es un vector independiente e y es una variable dependiente escalar.

Se definen varios tipos de regresión usando tres "ejes de clasificación":

* LINEALIDAD: Cuando un modelo parametrico es lineal en todos los parámetros del modelo, es decir que $f(x|\theta)$ es una combinación lineal de los parámetros θ con coeficientes que no dependen de los parámetros, se usa REGRESIÓN LINEAL (por ejemplo, regresión polinomial). Cuando los coeficientes de la combinación si dependen de los parámetros (por ejemplo, la función $f(x|\theta) = \theta_1 + \theta_2 . sen(\theta_3 . x)$) se usa REGRESIÓN NO LINEAL.

* COMPLEJIDAD DEL PROBLEMA: Un gran número de variables independientes aumenta la complejidad del problema. El caso de regresión más común es el que tiene una sola variable independiente.

Scanned with CamScanner

* COMPORTAMIENTO DEL ERROR: Las incertidumbres en los valores de las variables independientes y dependientes, y sus correlaciones, son el factor principal que determina qué método de regresión usar. En general, tomamos errores gaussianos.

- **REGRESIÓN LINEAL**

Considerando el modelo lineal:

$$y_i = \theta_0 + \theta_1 . x + \varepsilon_i$$

donde θ_0 y θ_1 son los coeficientes que describen la función de regresión (pendiente y ordenada al origen) y ε_i representa un término de error aditivo.

Suponiendo errores gaussianos para "x" y errores despreciables para "y", los parámetros del modelo lineal, θ_0 y θ_1, ajustado deben estimarse mediante los datos de la muestra usando método de cuadrados mínimos, tal que la suma de los cuadrados de las diferencias entre las observaciones y la recta de regresión sea mínima, es decir, se busca minimizar la función de costo:

$$S(\theta_0, \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \quad \text{dados} \quad \{(x_i, y_i)\}_{i=1}^{n}$$

|NOTA| Si los datos son multivariados (hay más de una variable independiente) se hace regresión multivariada, donde se ajusta un hiperplano en lugar de una recta. En este caso, la $f(\vec{x}, \theta)$ está dada por:

$$y_i = \theta_0 + \theta_1 . x_1 + \theta_2 . x_2 + \dots + \theta_k x_k + \varepsilon_i$$

con θ_i los parámetros de la regresión.

- **REGRESIÓN NO LINEAL**

En este caso el modelo es no lineal:

$$y = F(x, \theta) + \varepsilon$$

donde F es alguna función no lineal respecto a alguno de los parámetros θ. Este tipo de regresión tiene similitudes con la regresión lineal:

- Describen la relación x-y modelando una curva.
- Minimizan la función de costo.

la diferencia radica en la forma del modelo.

Scanned with CamScanner

Estos modelos pueden ser curvas cóncavas, convexas, exponenciales, sigmoidales, asintóticas, etc.

Al ser MODELOS NO LINEALES, los cuadrados mínimos se resuelven con métodos iterativos como Levenberg-Marquardt, Gauss-Newton o gradiente descendiente.

* LINEALIZACIÓN

Algunos problemas de regresión no lineal pueden linealizarse mediante una transformación del modelo. Por ejemplo,

dados a, b parámetros $y = a.e^{bx} \Rightarrow Ln(y) = Ln(a) + bx$

y el modelo es ahora lineal de $Ln(y)$ con respecto a x.

• PENALIZACIÓN DEL MODELO

Una vez planteado un modelo, vimos que minimizando cuadrados mínimos encontraremos el mejor ajuste (los mejores valores para los parámetros) del modelo dado.

Pero ¿Qué pasa si estamos planteando mal el modelo desde un inicio? Como, por ejemplo, la elección de un modelo muy complejo que derive en un sobreajuste de los datos. Para evitar esto, hay técnicas de penalización a la complejidad del modelo, que por lo general impone suavidad o limita el número de coeficientes en la regresión.

* Penalización del Likelihood => penaliza la función de probabilidad posterior.
* Regularización de Tikhonov => penaliza la suma de los cuadrados de los coeficientes de la regresión, tal que:

$|\theta| < s$ donde s controla la complejidad del modelo.

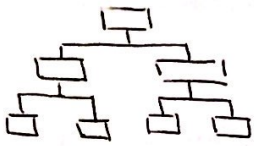
Scanned with CamScanner

CLASIFICACIÓN

RANDOM FOREST (técnica)
↙ BAGGING
↘ BOOSTING (como muestrear)

ÁRBOLES DE DESICIÓN

Es un procedimiento que divide la muestra multivariada de entrenamiento en clases mediante una división progresiva.



Dadas las features (dimensiones de un conjunto de datos, características), se elige la mejor y para ese feature se elige un valor de corte para separar la muestra según ese valor (generando dos ramas). Se repite el proceso para cada submuestra y se corta con algún criterio.

NODO: cuando se toma una desición.
RAMAS: bifurcaciones que dividen los datos respecto a un valor.
NODOS HOJA: ramas finales del árbol.

Hay muchos métodos para elegir los mejores features y los mejores valores de corte. Ej: greedy ⇒ prueba con todos los features y en cada feature con todos los valores para minimizar la función de costo:

Funciones de costo conocidas (medidas de impureza)
- ECM = $\sum (y-\hat{y})^2$
- Índice de Gini: $I = \sum P_k (1-P_k)$ → proporción de elementos de una clase después de una desición.

Existen técnicas de "podado" de los árboles, deteniendo su crecimiento. Si se lo deja crecer hasta que cada rama tenga submuestras puras, esas ramas tienen poco poder predictivo (reproduce muy bien los datos de entrenamiento) ⇒ sobreajuste.
El podado consiste en cortar ramas y evaluar el impacto en la performance. Si cambia poco se corta. Ej: penalización de complejidad o validación cruzada.

Pero es un método muy simple, ¿Cómo se aplica en la vida real?

- BAGGING → Es una forma de random forest (bootstrap + agregation)

La fiabilidad de una rama puede estimarse usando un conjunto de árboles construidos con réplicas bootstrap de los datos originales. Y promedia los resultados.

- RANDOM FOREST

Está formado por muchos árboles de desición individuales (formados con el mismo conjunto de entrenamiento). Cada árbol vota por una clase y la clase con más votos será la predicción del modelo.

Es muy eficaz, ya que las predicciones conjuntas son más precisas que las individuales.

Votación ⇒ clasificación.

Promediación ⇒ regresión.

Scanned with CamScanner

• BOOSTING (*)

Combina un gran número de clasificadores de predicción débil para obtener un procedimiento de clasificación fuerte.

Boosting está formado por un conjunto de árboles de desición individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

(*) Le da más peso a los casos que son más dificiles de predecir. se hacen iteraciones modificando los pesos en función del resultado del random forest en cada iteración (qué le costó más predecir).

CLASIFICADOR NAIVE BAYES

Dados N objetos con P propiedades c/u (features), $(x_1, \dots, x_N)$, que pertenecen a K grupos con distintas propiedades (clases), $(C_1, \dots, C_K)$.

Busco: dado un objeto ¿Qué clase le asigno en base a sus propiedades?

$$\Downarrow$$

$$P(C_K|\vec{x})$$

Usando el teorema de Bayes:

$$P(C_K|\vec{x}) = \frac{P(\vec{x}|C_K) \cdot P(C_K)}{P(\vec{x})}$$

$P(C_K)$ → prior → en este caso, es fácil de calcular.

$P(\vec{x})$ → evidencia → en este caso, es constante.

Con $P(\vec{x})$=cte queda:

$$P(C_K|\vec{x}) \propto P(\vec{x}|C_K) \cdot P(C_K) = \underbrace{P(\vec{x}, C_K)}_{\text{probabilidad condicional}} = P(x_1|x_2, \dots, x_N, C_K) \cdot P(x_2|x_3, \dots, x_N, C_K) \dots$$

Y suponiendo que cada propiedad es condicionalmente independiente de todas las otras propiedades:

$$P(x_i|x_{i+1}, \dots, x_N, C_K) = P(x_i, C_K) \Rightarrow \boxed{P(C_K|\vec{x}) \propto P(C_K) \cdot \prod_{i=1}^{N} P(x_i, C_K)}$$

Ejemplo: jugar → clases: jugar, no jugar (C_K)
atributos: soleado, nublado, lluvioso (x_i)
} busco: la probabilidad de jugar dado el clima.

PREGUNTAR POR ESTE ALGORITMO

Scanned with CamScanner

ALGORITMO DE EXPECTACIÓN/MAXIMIZACIÓN → charlar este con Marce

Es un método para encontrar el máximo del likelihood, $P(D|M)$ o $P(D,\vec{\theta})$, en modelos que tienen variables latentes (variables que participan de la formulación de un modelo, pero que no son directamente observadas sino calculadas de manera indirecta). → pedir a Marce ej de variables latentes

Consiste en un algoritmo iterativo de dos pasos: E-step y M-step.

① Comenzamos en algún punto del espacio de parámetros, asumiendo valores de los parámetros (estimación inicial), y queremos mejorar iterativamente esos valores.

② Con las estimaciones iniciales asumidas, se asignan pesos de acuerdo a los puntos de entrenamiento, estimando la distribución de la variable latente en función de los datos. |E-step|

③ Los pesos se usan para actualizar las estimaciones de los parámetros, usando el estimador según maximum likelihood pesado. |M-step|

④ Volvemos a ②.

Ejemplo

Necesito explicación pág 105.

Scanned with CamScanner

- MEZCLA DE GAUSSIANAS

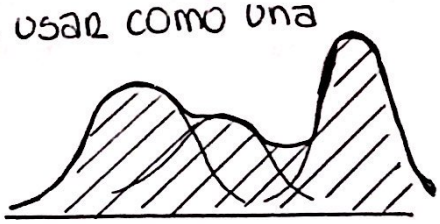
Supongamos que tenemos n objetos con p propiedades (features) c/u, que pertenecen a k grupos que se sabe que tienen distribución normal.

Nos preguntamos ¿Cómo son esas gaussianas?

La idea es que los parámetros que caracterizan esas gaussianas dependen de las probabilidades que tienen los n objetos $(x_1,\dots,x_n)=\vec{x}$ de pertenecer a los grupos.

Una mezcla de gaussianas está dada por: $f(x)=\sum_{i=1}^{k} a_i N(\vec{x};\mu_i,\sigma_i)$.

Si cada gaussiana tiene norma igual a 1 y la suma de los coeficientes $\sum_{i=1}^{k} a_i = 1$, entonces el área es igual a 1 y se puede usar como una densidad de probabilidad. Y para encontrar el conjunto de parámetros $(a_1,\dots,a_k)$, $(\mu_1,\dots,\mu_k)$ y $(\sigma_1,\dots,\sigma_k)$ se puede usar *Maximum Likelihood*:



$$\log\left(P(x|\vec{a},\vec{\mu},\vec{\sigma})\right)=\sum_{j=1}^{n}\log\left[\sum_{i=1}^{k} a_i . N(x;\vec{\mu},\vec{\sigma})\right]$$

pero para resolver esto hay que encontrar todos los parámetros de todas las gaussianas y es muy difícil. Pero se puede usar EXPECTATION/MAXIMI-ZATION.

Se define una variable latente $\vec{z}$ tal que $\vec{z}=(0,\dots,z_k=1,\dots,0)$ y cada vector $\vec{z}$ tiene k estados posibles, según donde esté el 1.

¿y cómo termina?

Scanned with CamScanner

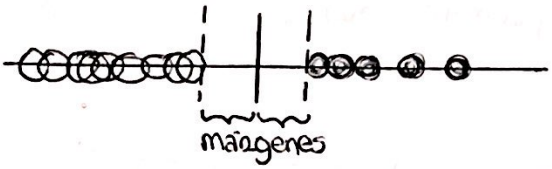
MARGIN CLASSIFIERS

Supongamos datos linealmente separables, es decir que existe al menos un hiperplano que separa dos categorías, llamado SUPERFICIE DE DECISIÓN.
Queremos un algoritmo que aprenda con un conjunto de entrenamiento y estime la clasificación de un conjunto de validación. El modelo es:

$$y(\vec{x}) = \vec{\omega}^T \phi(\vec{x}) + \vec{b} \quad \text{con} \quad \left.\begin{array}{l} \vec{\omega}^T\text{: orientación del hiperplano.} \\ \vec{b}\text{: sesgo.} \end{array}\right\} \text{parámetros del modelo}$$

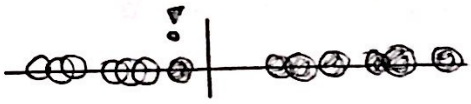
Las arquitecturas que resuelven estos problemas usan el concepto de MARGEN: distancia mínima entre la superficie de desición y cualquiera de los elementos del conjunto de entrenamiento. Los algoritmos de "clasificación con margen" consisten en encontrar la superficie que maximiza el margen (algoritmos de optimización).

- MAXIMUM MARGIN CLASSIFIERS (clasificador de máximo margen)



Cuando buscamos maximizar los márgenes tenemos un clasificador de máximo margen.

Podemos permitir elementos mal clasificados para mejorar la capacidad de predicción, ya que debe ser un compromiso entre sesgo y varianza.



El margen se calcula como la distancia perpendicular de la superficie de decisión a los puntos más cercanos, llamada "support vectors" (vectores de soporte).
↳ solo estos puntos son relevantes para definir la superficie a partir de los datos de entrenamiento mediante un procedimiento de optimización que maximiza el margen (mover los otros puntos no influye en los márgenes).

$\vec{\omega}^T$: es siempre perpendicular al plano y da la orientación del plano.
$\vec{b}$: es la ordenada al origen (a donde se mueve el plano).
$\phi(\vec{x})$: es el kernel (la posición de cada punto luego de aplicar el kernel).
$y(\vec{x})$: es la ec. de un plano en "n" dimensiones (me da la distancia de cada punto a la hipersuperficie (plano en 3D).

Dadas dos clases $t_n = 1$ y $t_n = -1$ entonces $\begin{cases} t_n = 1 \Rightarrow y(\vec{x}) > 0 \\ t_n = -1 \Rightarrow y(\vec{x}) < 0 \end{cases}$

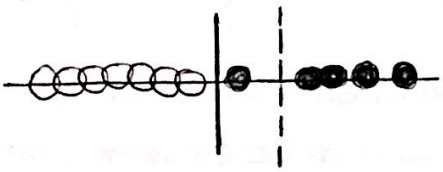
$\Rightarrow$ se busca minimizar $|\vec{\omega}|$ (maximizar $\frac{1}{|\vec{\omega}|}$)

Scanned with CamScanner

• **SOFT MARGIN CLASSIFIERS** (clasificador de márgenes suave)

Si los datos tienen outliers, el SMC permite clasificaciones incorrectas con umbrales "blandos" con el fin de elegir el mejor plano en un compromiso entre sesgo y varianza.

Ejemplo:



#Línea sólida => muy sensible a los datos de entrenamiento, ajusta fluctuaciones y por lo tanto tiene alta varianza.

#Línea dash => permite clasificaciones incorrectas y es menos sensible a los datos de entrenamiento, por lo tanto, baja varianza.

¿Cómo eligo los márgenes? ¿Cuál de las dos líneas es mejor?

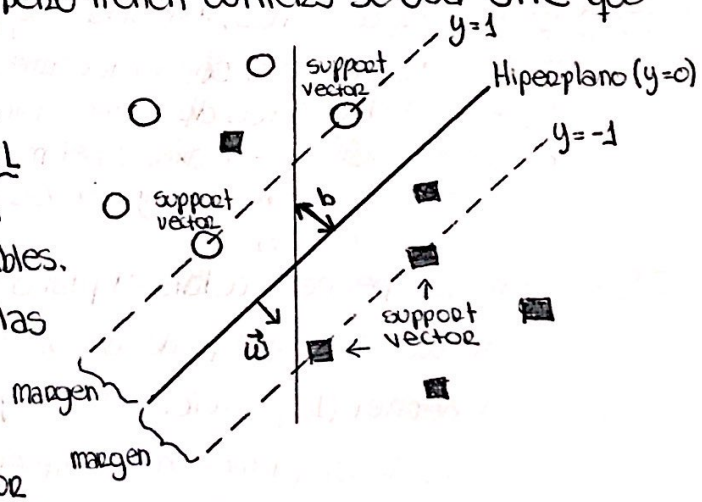
Para esto se usa VALIDACIÓN CRUZADA ⇨ Se establece un error que sea la suma de la diferencia entre las etiquetas asignadas y medidas => hago muchos margenes y me quedo con aquellos que tengan menos error en el test sample.

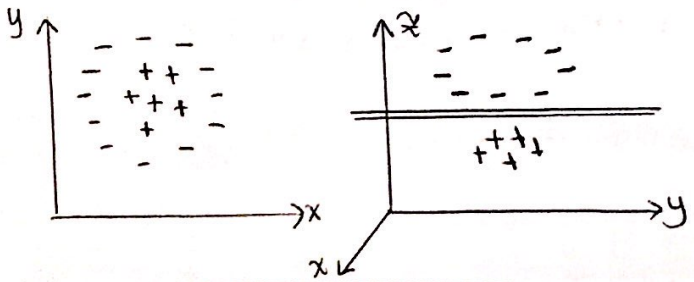
• **SUPPORT VECTOR MACHINES** (SVM)

Es una generalización de la idea de los clasificadores de margen. Dado un conjunto de entrenamiento, SVM busca encontrar un hiperplano que clasifique nuevos datos.

(en 1D es un punto, en 2D una recta, en 3D un plano).

* Si los datos son linealmente separables y no tienen outliers se usa HMC que también se llama Hard Margin SVM.
* Si los datos son linealmente separables pero tienen outliers se usa SMC que también se denomina Soft Margin SVM.
* Si los datos no son linealmente separables (no se puede trazar una línea) se usa KERNEL TRICK, que lleva los datos a una dimensión más alta para hacerlos linealmente separables. Permiten tratar problemas no lineales con las herramientas de problemas lineales. Se aplica un kernel $\phi(\vec{x})$ a los datos, proyectando los datos a un espacio mayor dimensión en donde se puede encontrar un hiperplano que los separe.





Scanned with CamScanner

REGRESIÓN LOGÍSTICA

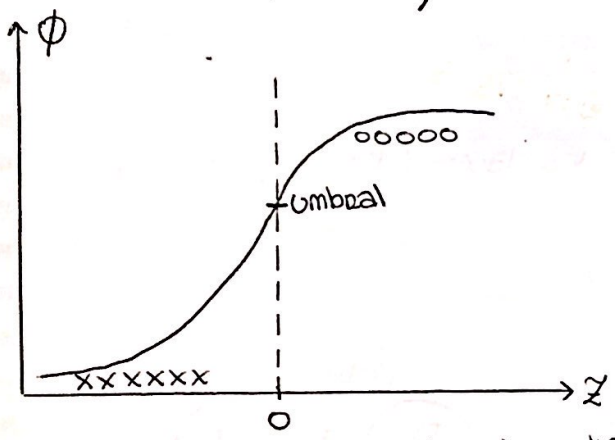
Sirve para clasificar dos clases linealmente separables. Busca una relación lineal para el Likelihood de una de las clases con el objetivo de modelar la probabilidad de que exista una determinada clase.

Utiliza una función logística para modelar una variable dependiente binaria. Dadas las clases 0 y 1, la función logística me dice: tomá un x y la $f(x)$ es la probabilidad de pertenecer a la clase 1.

$$\mathcal{L} = \frac{p(X)}{1-p(X)}$$ con $p(X)$ = proporción de elementos de una clase.
X = datos.

Conviene escribirlo como: $\log\left(\underbrace{\frac{p(X)}{1-p(X)}}_{\text{likelihood de una de las clases.}}\right) = ax+b \Rightarrow p(X) = \frac{1}{1+e^{-a\cdot x-b}}$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \log\left(\frac{P(C=0/X=x)}{P(C=1/X=x)}\right)$$



$$\phi(z) = \frac{1}{1+e^{-z}}$$ FUNCIÓN LOGÍSTICA
↓
Varía según a y b.

Defino un umbral ⇒ hacia arriba son una clase y hacia abajo otra cambiando a y b (cambiando la función logística).

El modelo tiene una variable dependiente con dos valores posibles: 0 y 1. El logaritmo de la probabilidad de una clase es una combinación lineal. La probabilidad del valor etiquetado como 1 varía entre 0 y 1. La función logística convierte los logaritmos en probabilidades.

Con el conjunto de entrenamiento se construye la función logística variando a y b.

- **FUNCIÓN DE COSTO USADA EN RL**

Para un clasificador hay que tener en cuenta
- El modelo
- La función de costo
- La estrategia para minimizar la función de costo.

En general venimos usando como función de costo "cuadrados mínimos" (por ejemplo, en regresión lineal). Si usamos cuadrados mínimos en regresión logística obtenemos una superficie que no es convexa (y necesitamos que sea convexa para tener un mínimo bien definido).

Scanned with CamScanner

Entonces, para cuantificar la diferencia entre el modelo y los datos, vamos a proponer una función de costo, J, que penalice las etiquetas "mal puestas".

$$J(\vec{x},y;\theta) = \frac{1}{m}\sum_{i=1}^{m} costo(\theta^i, y^i) = \frac{1}{m}\sum_{i=1}^{m}\left[-y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))\right]$$

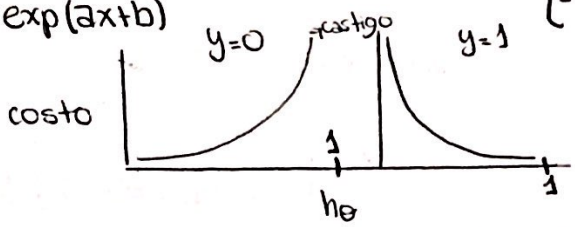
con

m: observaciones.

h_θ: $h_\theta(x) = \dfrac{1}{1+\exp(ax+b)}$

y: clase

$$costo(\theta,y) = \begin{cases} -\log(h_\theta(x)) & y=1 \\ -\log(1-h_\theta(x)) & y=0 \end{cases}$$



Se busca minimizar la función de costo total $J(\vec{x},y;\theta)$ para encontrar los valores óptimos de $\vec{\theta}$: a y b con el conjunto de entrenamiento.

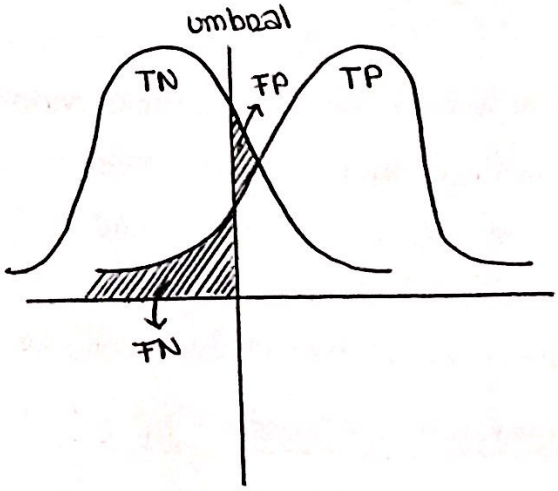
↳ Método más usado: gradiente descendiente.

MÉTRICAS

Supongamos que los datos se separan en dos clases (dos etiquetas): positivo y negativo.

MATRIZ DE CONFUSIÓN

	valores reales	
valores predichos por el modelo	TP verdaderos positivos	FP falsos positivos
	FN falsos negativos	TN verdaderos negativos



La matriz de confusión describe el rendimiento del modelo en el conjunto de validación. Con el conjunto de entrenamiento se determina el umbral y luego se usa el conjunto de validación para determinar:

- TP: dato positivo clasificado como positivo.
- FP: dato negativo clasificado como positivo.
- TN: dato negativo clasificado como negativo.
- FN: dato positivo clasificado como negativo.

Scanned with CamScanner

Entre más diagonal sea la matriz, mejor clasifica el modelo, es decir, mejor es la capacidad de predicción del modelo.

TIPOS DE MÉTRICAS

- **TPR (tasa de positivos verdaderos)**

Fracción de veces que valores positivos reales son identificados como positivos.

$$TPR = \frac{TP}{TP + FN}$$

<recall> o <eficiencia> Esto mede un valor.

- **FPR (tasa de falsos positivos)**

Fracción de veces que valores negativos reales son identificados equivocadamente como positivos.

$$FPR = \frac{FP}{FP + TN}$$

- **Precisión**

Fracción de veces que los que son identificados como positivos son realmente positivos.

$$Precisión = \frac{TP}{TP + FP}$$

<completitud>

- **FOR (false omission rate)**

Fracción de los que son identificados como negativos que son positivos (fueron mal identificados).

$$FOR = \frac{FN}{FN + TN}$$

- **Accuracy**

Fracción de toda la muestra que está bien clasificada.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

CURVAS

Cada algoritmo debe ser probado con más de una métrica, para ver qué algoritmo es mejor.

Cada métrica es buena para alguna cosa pero mala para otra. Las métricas más usadas:

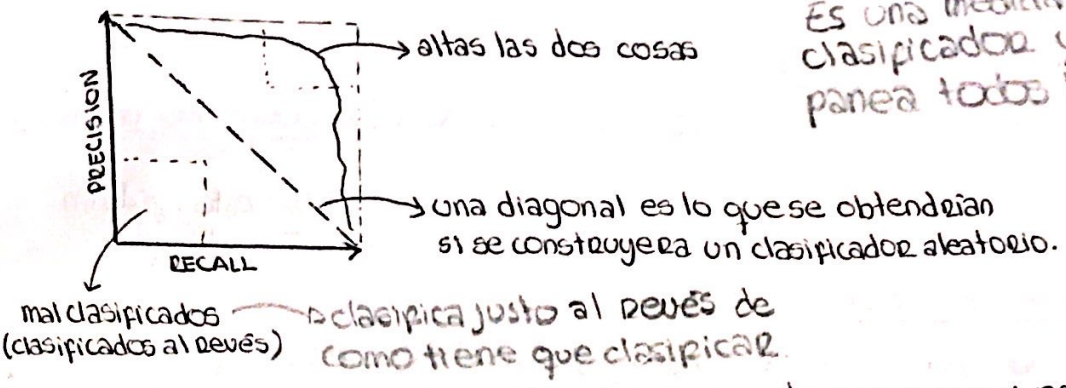
- Curva precisión-recall. → muestra desbalanceada.
- Curva ROC. → muestra balanceada.
- F1-Score. → muestra desbalanceada.

Hay un compromiso entre las métricas: cuando modifico el valor umbral puedo optimizar uno pero empeoro otros.

Scanned with CamScanner

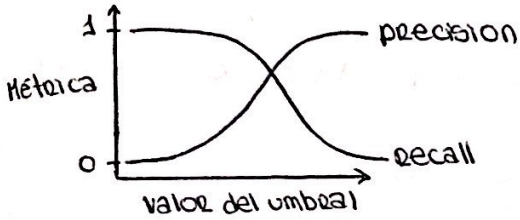
○ CURVA PRECISION-RECALL

Es la completitud vs la eficiencia. Quiero que ambas sean altas. → que tan buenos es.



Es una medida sobre el clasificador ya que parea todos los umbrale

Para construir la curva se calculan las curvas de precision y recall para el mismo modelo y mismo conjunto de validación tomando distintos valores umbrales. En gral, se comparan varias curvas de diferentes modelos. (*)



Problema: si aumento uno, disminuye el otro.

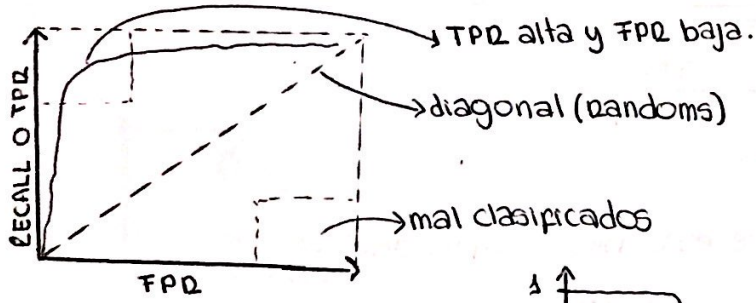
Aumento el umbral ⇒ aumento precisión

Disminuyo el umbral ⇒ aumento recall.

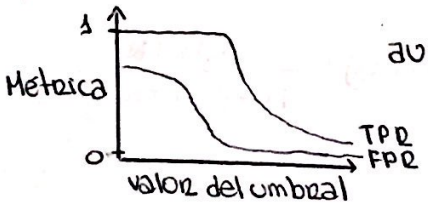
(*) Al mover el umbral tengo diferentes puntos en estos gráficos.

○ CURVA ROC

Es la eficiencia en función de la tasa de falsos positivos.



El área bajo la curva ROC es una métrica también: si me da 1 la curva es perfecta, si me da 0 estoy en la diagonal (predicción aleatoria).



aumenta el umbral
↳ mejora FPR (pocos FP).

○ F1-SCORE

Buscar sobre esto

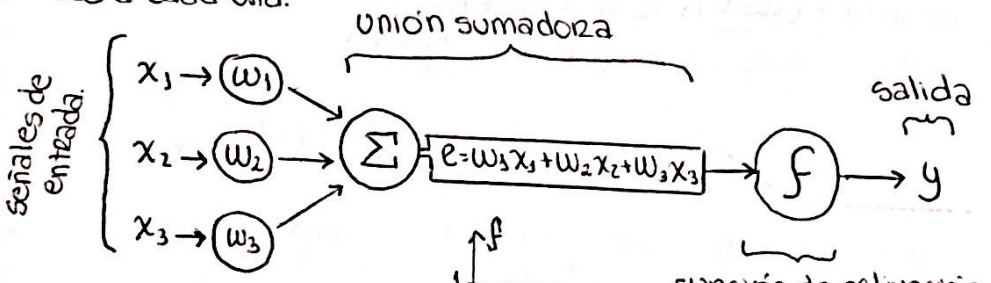
Scanned with CamScanner

DEEP LEARNING ⇒ redes neuronales con muchas capas.

PERCEPTRON

Un perceptron es una neurona artificial (clasificador lineal o binario) y, por lo tanto, una red neuronal (la más simple) de una sola capa.

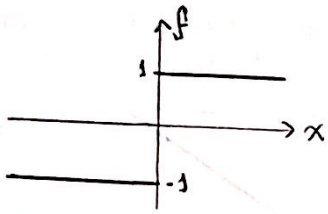
Consiste en un hiperplano que separa el espacio en dos regiones y asigna una clase a cada una.



En este caso, la f es de tipo signo:

$$f = \begin{cases} 1 & \text{si } \sum w_i x_i > \theta \\ -1 & \text{si } \sum w_i x_i < \theta \end{cases}$$

con 1 o -1 la salida.



Funcionamiento:

- Los datos de entrada se multiplican por los pesos y el resultado es un valor.
- Ese valor puede ser positivo o negativo.
- La neurona se activa si el valor es positivo. Solo se activa si el peso calculado de los datos de entrada supera un umbral determinado, θ.
- El resultado predicho se compara con el resultado conocido. En caso de diferencia, el error se retropropaga para permitir ajustar los pesos.

* Enseñar al perceptron quiere decir calcular los pesos. El problema es encontrar los pesos adecuados que reproducen los outputs del cjto de entrenamiento.

* La salida es el resultado de aplicar una función f a la activación, dando la función de activación. La salida se denomina actividad de la neurona.

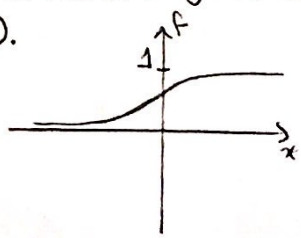
NEURONA SIMPLE

Es un modelo similar al de regresión logística o al perceptron, solo que puede tener cualquier función de activación f.

* FUNCIONES DE ACTIVACIÓN

La función de activación se encarga de devolver una salida a partir de un valor de entrada.

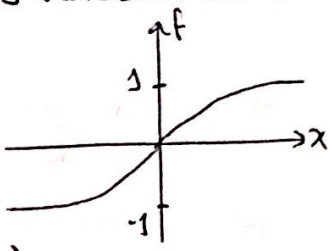
- Sigmoide: Transforma los valores introducidos a una escala (0,1) donde los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a 0.



$$f(x) = \frac{1}{1 - e^{-x}}$$

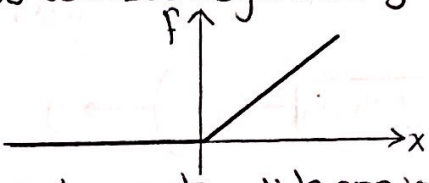
Scanned with CamScanner

• Tangente hiperbólica: Transforma los valores de entrada a una escala (-1,1), donde los valores altos van asintóticamente a 1 y los bajos a -1.



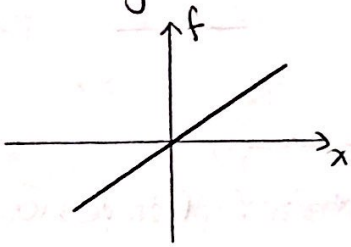
$$f(x) = \frac{\cdot}{1+e^{-2x}}$$

• Rectified Lineal Unit (ReLU): Transforma los valores de entrada anulando los valores negativos y dejando igual a los positivos.



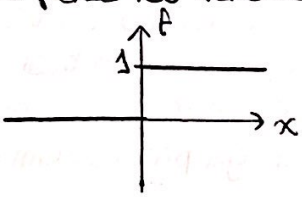
$$f(x) = \max(0,x) = \begin{cases} 0 & x<0 \\ x & x\geq 0 \end{cases}$$

• Lineal: Los valores de salida son iguales a los de entrada.



$$f(x) = x$$

• Escalón: La salida es cero para los valores de entrada negativos y uno para los positivos.



$$f(x) = \begin{cases} 0 & x<0 \\ 1 & x\geq 0 \end{cases}$$

NOTA Las redes más complejas se construyen a partir de un conjunto de neuronas simples.

ARQUITECTURA

Una neurona pondera los datos de entrada y posteriormente modifica este resultado usando una función de activación.

① Una neurona simple tiene un conjunto de valores de entrada (inputs), $\vec{x} = (x_1, \dots, x_n)$, y un valor de salida (output) y. Asociado a cada input hay un peso w_i y puede haber un valor adicional w_0 (bias).

② Se calcula la activación sobre todos los inputs: $a = \sum w_i x_i$ → inputs.

③ El output es el resultado de aplicar una función f a la activación: función de activación.

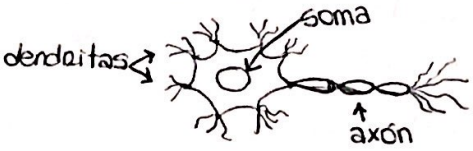
④ La salida se denomina actividad de la neurona.

Scanned with CamScanner

REDES NEURONALES

Sirven para tareas de reconocimiento de patrones, control de dispositivos, clasificación de objetos, etc.

Los seres vivos contamos con redes neuronales biológicas, cada neurona biológica posee un canal de entrada de información (dendritas), un órgano de cómputo (soma) y un canal de salida (axón):



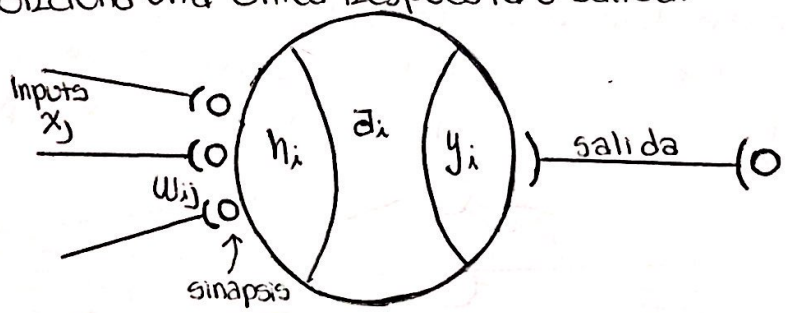
Cuando una neurona se conecta con otra se da la sinápsis. La neurona que envía la info. se llama presináptica y la que recibe postsináptica.

En el cerebro, el axón de una neurona entrega su información como señal de entrada a la dendrita de otra neurona, la cual deriva la información al soma para hacer la sinapsis. Este proceso se repite muchas veces, formando una red.

Las redes neuronales artificiales se inspiran en las biológicas. Son el resultado de interconectar multiples neuronas simples.

- MODELO GENÉRICO DE UNA NEURONA ARTIFICIAL

Una neurona es un procesador que a partir de un vector de entrada (que viene del input inicial exterior o de la salida de un conjunto de neuronas), proporciona una única respuesta o salida.



Los elementos que conforman dicha neurona son:

* Entradas: x_j.

* Pesos sinápticos: w_{ij}. Representan la intensidad de interacción entre cada neurona presináptica j y la neurona postsináptica i.

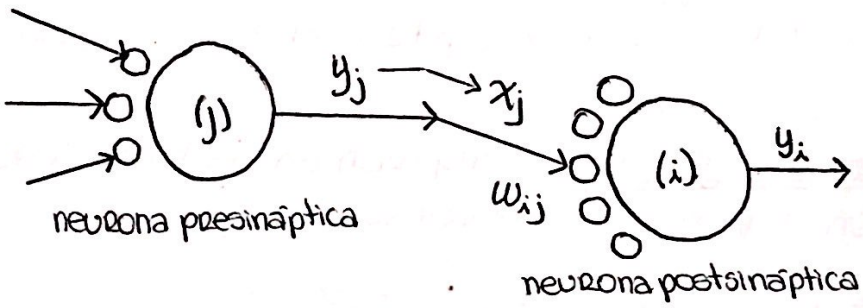
* Regla de aprendizaje/propagación: $h_i = \sigma(w_{ij}, x_j(t))$. Proporciona el valor del potencial postsináptico, $h_i(t)$, de la neurona i en función de sus pesos y sus entradas, el cual representa la manera en que los pesos cambian con el tiempo.

Scanned with CamScanner

Dada una entrada positiva, si el peso es positivo tenderá a excitar a la neurona postsináptica y si el peso es negativo tenderá a inhibirla.

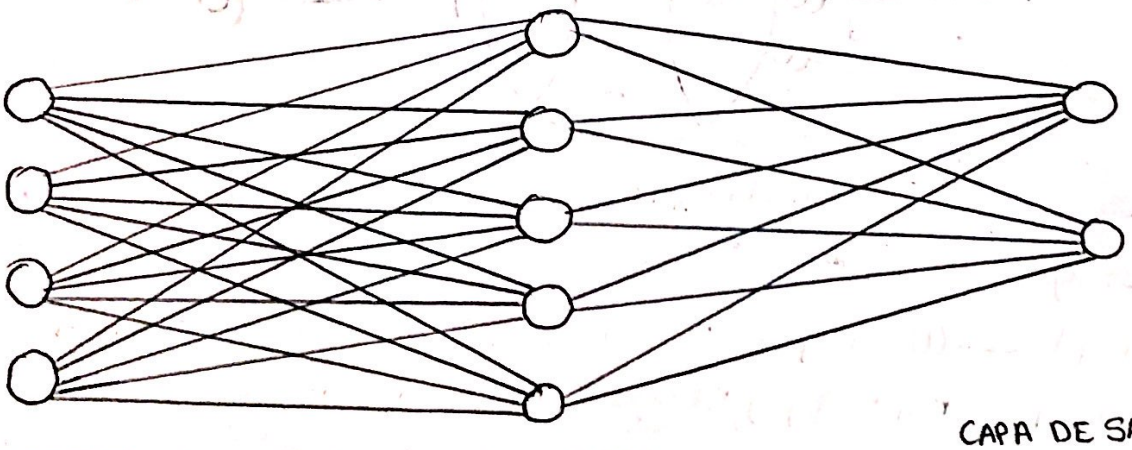
* Regla/función de activación: $f_i(a_i(t-1), h_i(t))$. Proporciona el estado de activación actual, $a_i(t)$, de la neurona i en función de su estado anterior, $a_i(t-1)$ y de $h_i(t)$: $a_i(t) = f_i(a_i(t-1), h_i(t))$.
Hay modelos donde el estado actual de la neurona no depende del anterior y, en ese caso: $a_i(t) = f_i(h_i(t))$.

- **INTERCONEXIÓN ENTRE NEURONAS**



- **ARQUITECTURA DE REDES NEURONALES**

En una red las conexiones sinápticas son direccionales, es decir, la información solo puede fluir en un sentido. Las neuronas se agrupan en unidades estructurales denominadas capas. Dentro de una capa las neuronas suelen ser del mismo tipo. Hay tres tipos de capas:



CAPA DE ENTRADA
compuesta por neuronas que reciben datos o señales procedentes del entorno.

CAPA OCULTA
no tiene una conexión directa con el entorno.

CAPA DE SALIDA
aquella cuyas neuronas proporcionan la respuesta de la red neuronal.

Scanned with CamScanner

* Si tiene una única capa oculta se denomina <u>red neuronal simple</u>.
* Para lograr mejores resultados de clasificación se agregan más neuronas y más capas. (profundizamos la red).
* Al conectar y entrenar las neuronas, cada capa es resultado de una combinación no lineal de las capas anteriores.
* La cantidad de parámetros que hay en una neurona es el número de flechas que llega a ella + un bias que tiene la neurona.
* Difiere de la regresión logística porque:
  - ✓ Pueden separar clases que no son linealmente separables.
  - ✓ Es un modelo no paramétrico (RL es paramétrico).
  - ✓ Sus parámetros son muy difíciles de interpretar.

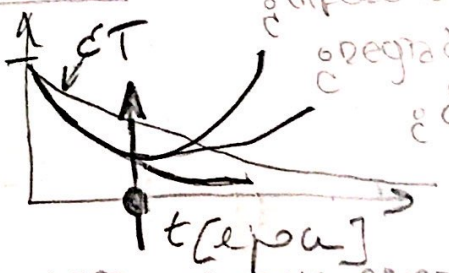
• <u>BACKPROPAGATION</u>

¿Cómo optimizo la red? es decir ¿cómo busco el mínimo en una red?

La técnica de backpropagation propaga los errores para atrás.

El propósito de entrenar la red es encontrar pesos que reduzcan el error entre el conjunto de salida y los datos de prueba. Esto se hace minimizando el error de cuadrados mínimos, con métodos como gradiente descendiente o Levenberg Marquardt.

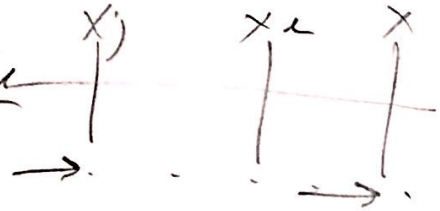
$$\frac{\partial \psi^2}{\partial w} = 0$$



¿diferenciando?
¿regla de la cadena?
¿δ?
Curva de Aprendizaje

VER ARQ. ULTIMA PARTE MARIANO

$$\bar{y} = \sum_i g(w_{ij} x_j)$$



$$= \sum_i g\left(w_{ij} \sum_k g(w_{ke} x_e)\right)$$

$$\bar{y} = f'(x) \mid (y - \bar{y})$$

$$\chi^2 (w's)$$

Scanned with CamScanner

PREGUNTAS

• Marce.
* Mariano.

• Charlar sobre el valor p. (12)
* ¿Qué saber de NMF? (21)
* Explicación del IsoMap (21b)
* clustering jerárquico (23b)
• Clasificador Naive Bayes (25b).
• Maximization/Exp. (26)
¿donde entra kernel trick?
• Importance sampling: ¿Qué pasa con la f(x) en la cuenta? (17)
• Criterios de información
LASSO y Ridge.

HIPERPARÁMETROS / PARÁMETROS

Primero se define la arquitectura recorriendo el espacio de hiperparámetros con técnicas como cross-validation viendo en el espacio de hiperparámetros dónde se minimiza la función de costo.
Una vez definida la arquitectura del modelo se definen los parámetros.

Ej: hiperparámetros → cantidad de capas y neuronas } redes neuronales.
parámetros → pesos

Scanned with CamScanner

TEMAS DE NOTEBOOK DE MARIANO MÁS DESARROLLADOS

Validación cruzada

El enfoque más simple para la validación cruzada es dividir los datos de entrenamiento en tres partes: el conjunto de entrenamiento, el conjunto de validación cruzada y el conjunto de prueba. Como regla general, el conjunto de entrenamiento debe comprender del 50 al 70 % de los datos de entrenamiento originales, mientras que el resto se divide por igual en el conjunto de validación cruzada y el conjunto de prueba.

El conjunto de entrenamiento se usa para determinar los parámetros de un modelo dado (es decir, los valores óptimos de θ). Usando el conjunto de entrenamiento, evaluamos el error de entrenamiento. El conjunto de validación cruzada se utiliza para evaluar el error de validación cruzada del modelo. Debido a que este conjunto de validación cruzada no se usó para construir el ajuste, el error de validación cruzada será grande para un modelo de alto sesgo (sobreajuste), y representa mejor la verdadera bondad del ajuste del modelo. Teniendo esto en cuenta, es probable que el modelo que minimice este error de validación cruzada sea el mejor modelo en la práctica. Una vez que se determina este modelo, el error de prueba se evalúa utilizando el conjunto de prueba. Este error de prueba da una estimación de la fiabilidad del modelo para un conjunto de datos no etiquetados.

¿Por qué necesitamos un conjunto de prueba y un conjunto de validación cruzada? En cierto sentido, así como los parámetros (en este caso, θ) se aprenden del conjunto de entrenamiento, los llamados hiperparámetros, aquellos parámetros que describen la complejidad del modelo, se aprenden del conjunto conjunto de validación cruzada. De la misma manera que los parámetros pueden sobreajustarse a los datos de entrenamiento, los hiperparámetros pueden sobreajustarse a los datos de validación cruzada, y el error de validación cruzada da una estimación demasiado optimista del rendimiento del modelo en un conjunto de datos sin etiquetar. El error de prueba es una mejor representación del error esperado para un nuevo conjunto de datos.

Curvas de aprendizaje

Una pregunta que la validación cruzada no aborda directamente es cómo mejorar un modelo que no está dando resultados satisfactorios. Hay varias posibilidades:

- Obtener más datos de entrenamiento. A menudo, usar más datos para entrenar un modelo puede generar mejores resultados.
- Usa un modelo más/menos complicado. Como vimos anteriormente, la complejidad de un modelo debe elegirse como un equilibrio entre el sesgo y la varianza.

Scanned with CamScanner

- Usar más/menos regularización. En general, aumentar la regularización tiene un efecto similar a disminuir la complejidad del modelo.

A medida que aumentamos el tamaño del conjunto de entrenamiento, el error de validación cruzada disminuye. La razón de esto es fácil de ver: un conjunto de entrenamiento más pequeño conduce a un ajuste excesivo del modelo, lo que significa que el modelo es menos representativo de los datos de validación cruzada.

A medida que crece el conjunto de entrenamiento, se reduce el sobreajuste y disminuye el error de validación cruzada. Nuevamente, además de las fluctuaciones aleatorias, y siempre que el conjunto de entrenamiento y el conjunto de validación cruzada sean estadísticamente similares, esperamos que el error de validación cruzada siempre disminuya a medida que crece el tamaño del conjunto de entrenamiento.

El error de entrenamiento es en todas partes menor o igual que el error de validación cruzada, hasta pequeñas fluctuaciones estadísticas.

Otras técnicas de validación-cruzada

Existen numerosas técnicas de validación cruzada disponibles que son adecuadas para diferentes situaciones.

Doble validación cruzada

Dividimos los datos en un conjunto de entrenamiento d1 , un conjunto de validación cruzada d2 y un conjunto de prueba d0 . Nuestras pruebas simples involucraron entrenar el modelo en d1 y validar de forma cruzada el modelo en d2 . En la doble validación cruzada, este proceso se repite, entrenando el modelo en d0 y validando de forma cruzada el modelo en d1 . El error de entrenamiento y el error de validación cruzada se calculan a partir de la media de los errores en cada pliegue. Esto conduce a una determinación más sólida del error de validación cruzada para conjuntos de datos más pequeños.

Validación cruzada de pliegues K

Una generalización de la validación cruzada doble es la validación cruzada de K veces. Aquí dividimos los datos en K + 1 conjuntos: el conjunto de prueba d0 y los conjuntos de validación cruzada d1,d2,...,dK. Entrenamos K modelos diferentes, cada vez dejando fuera un solo subconjunto para medir el error de validación cruzada. El error de entrenamiento final y el error de validación cruzada se pueden calcular utilizando la media o la mediana del conjunto de resultados. La mediana

Scanned with CamScanner

puede ser una mejor estadística que la media en los casos en que los subconjuntos contienen pocos puntos.

Series temporales

Una **serie temporal** o cronológica es una sucesión de datos medidos en determinados momentos y ordenados cronológicamente. Los datos pueden estar espaciados a intervalos iguales (como la temperatura en un observatorio meteorológico en días sucesivos al mediodía) o desiguales (como el peso de una persona en sucesivas mediciones en el consultorio médico, la farmacia, etc.). Para el análisis de las series temporales se usan métodos que ayudan a interpretarlas y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series y que permiten en diferente medida y con distinta confianza extrapolar o interpolar los datos y así predecir el comportamiento de la serie en momentos no observados, sean en el futuro (extrapolación pronóstica), en el pasado (extrapolación retrógrada) o en momentos intermedios (interpolación). Uno de los usos más habituales de las series de datos temporales es su análisis para predicción y pronóstico.

A diferencia de los problemas de regresión en los que las distintas medidas de "y" suelen tratarse como variables aleatorias independientes, en tales modelos el valor de "yi+1" depende directamente del valor precedente "yi".

Las principales tareas del análisis de series temporales son:

(1) caracterizar la supuesta correlación temporal entre diferentes valores de "y".
(2) pronosticar (predecir) valores futuros de y.

Dada una serie de tiempo, a menudo primero queremos determinar si hemos detectado variabilidad. Esto es equivalente a preguntar si los datos son consistentes con la hipótesis nula descrita por un modelo que consiste en una señal constante más el ruido de medición.

Herramientas para el análisis:
Primero revisaremos los métodos presentados anteriormente (principalmente aplicables al análisis en el dominio del tiempo) y discutiremos la estimación de parámetros, la selección de modelos y la clasificación en el contexto del análisis de series de tiempo. Luego, ampliaremos este conjunto de herramientas mediante la introducción de herramientas para el análisis en el dominio de la frecuencia, como el análisis de Fourier.

Scanned with CamScanner

Estimación de parámetros, selección de modelos y clasificación por tiempo

En general, ajustamos un modelo a un conjunto de N puntos de datos con errores conocidos para "y":

$$y_j(t_j) = \sum_{m=1}^{M} \beta_m T_m(t_j|\theta_m) + \epsilon_j$$

donde las funciones ***Tm(t|θm)*** no necesitan ser periódicas, ni los tiempos ***tj*** necesitan ser muestreados uniformemente. Como antes, el vector ***θm*** contiene parámetros del modelo que describen cada ***Tm(t)*** (aquí usamos el símbolo | para significar "parámetros dados θm" y no en el sentido de un pdf condicional).

Los modelos deterministas comunes para el proceso subyacente que genera datos incluyen ***T(t)=sin(ωt)*** y ***T(t)=exp(−αt)***, donde la frecuencia ***ω*** y la tasa de decaimiento ***α*** son parámetros del modelo que se estiman a partir de los datos. Otro modelo importante es la llamada "señal chirp", ***T(t)=sin(ϕ+ωt+αt2)***. En la ec. ***ϵ*** significa ruido, que normalmente se describe mediante errores gaussianos

Análisis de Fourier

La transformada de Fourier y su transformada inversa:

$$H(f) = \int_{-\infty}^{\infty} h(t) exp(-i2\pi ft) dt$$

$$h(t) = \int_{-\infty}^{\infty} H(f) exp(i2\pi ft) df$$

donde t es el tiempo y f es la frecuencia.

Una convolución de dos funciones a(t) y b(t) viene dada por:

$$(a \star b)(t) = \int_{-\infty}^{\infty} a(t')b(t-t')dt'.$$

Por ejemplo, en imágenes astronómicas, la distribución de intensidad real en el cielo se convoluciona con la visión atmosférica para imágenes basadas en tierra, o el patrón de difracción del telescopio para imágenes basadas en el espacio. En la ecuación anterior, la función ***a*** puede considerarse como el "patrón convolutivo" del aparato de medición, y la función ***b*** es la señal.

Scanned with CamScanner

En la práctica, medimos la versión convolucionada (o suavizada) de nuestra señal, ***a⋆b(t)***, y buscamos descubrir la señal original ***b*** usando la presumiblemente conocida ***a***: deconvolucionamos.

Transformada de Fourier discreta

En la práctica, los datos siempre se muestrean discretamente. Cuando el espaciado del intervalo de tiempo es constante, la transformada discreta de Fourier es una herramienta poderosa. Supongamos que tenemos una función real continua ***h(t)*** que se muestrea en N intervalos iguales ***hj=h(tj)*** con ***tj≡t0+Δtj,j=0,...,(N−1)***, donde el intervalo de muestreo t y la duración de la toma de datos T están relacionados a través de ***T=NΔt*** (el agrupamiento podría haber sido realizado por el aparato de medición, por ejemplo, imágenes CCD, o durante el análisis de datos). La transformada discreta de Fourier y su transformada inversa:

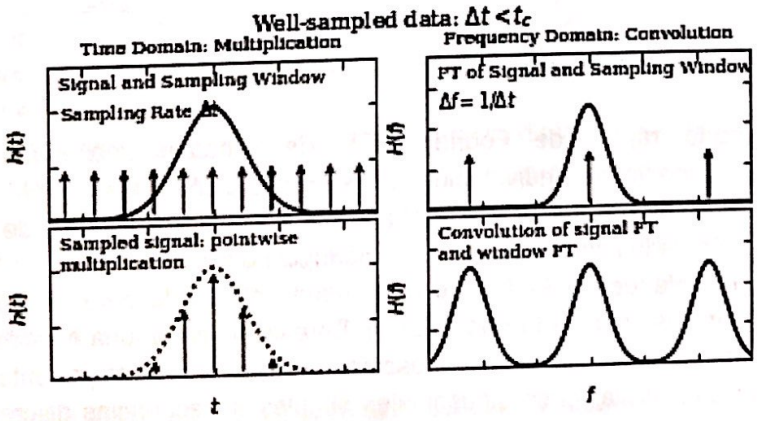
$$H_k = \sum_{j-0}^{N-1} h_j exp[-i2\pi jk/N]$$

$$h_j = \frac{1}{N}\sum_{k=0}^{N-1} H_k exp[i2\pi jk/N]$$

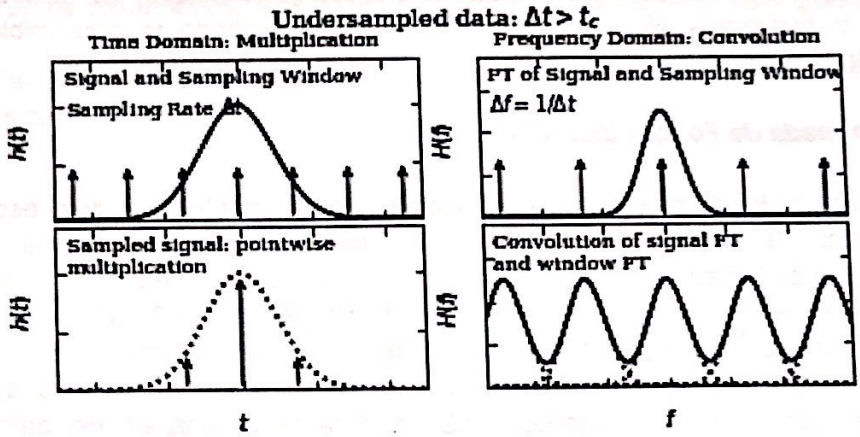
Tratamiento estadísticos de datos.

La función ventana → (Es lo que hace Cami).

La figura siguiente muestra la relación entre el muestreo y la función ventana: la función ventana se puede expresar en el dominio del tiempo como la suma de las funciones delta colocadas en los tiempos de observación muestreados. <u>En este caso las observaciones están espaciadas regularmente</u>.
La transformada de Fourier de un conjunto de funciones delta con espaciado Δt es otro conjunto de funciones delta con espaciado 1/Δt.

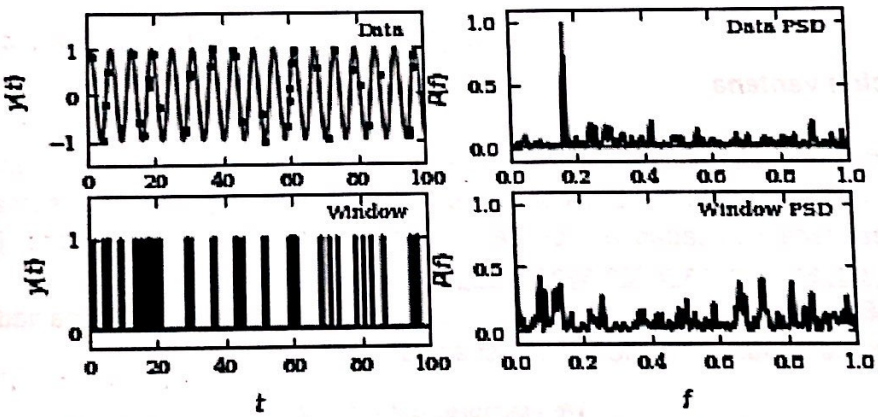


Scanned with CamScanner



Cuando los datos se muestrean de manera no uniforme, si bien la ventana de muestreo es la suma de las funciones delta, pero debido a que las funciones delta no están espaciadas regularmente, la transformada de Fourier es una función de f más complicada y, en general, compleja.

En la siguiente figura se muestra un ejemplo de una ventana de muestreo irregular: aquí la verdadera transformada de Fourier de los datos sinusoidales es un pico localizado.



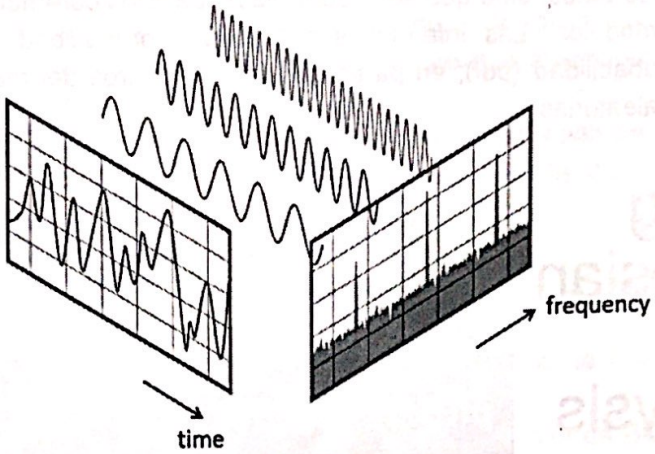
La transformada rápida de Fourier

La transformada rápida de Fourier (FFT) descompone una señal en sus componentes espectrales individuales y así proporciona información sobre su composición. Los FFT se utilizan para el análisis de errores, el control de calidad y la monitorización de las condiciones de las máquinas o sistemas.

Estrictamente hablando, la FFT es un algoritmo optimizado para implementar la "Transformación Discreta de Fourier" (DFT). En este proceso, una sección limitada en el tiempo de una señal se descompone en sus componentes. Estos componentes son oscilaciones sinusoidales simples a frecuencias discretas, cuya

Scanned with CamScanner

amplitud y fase están determinadas. El FFT permite así la vista de una señal en el dominio de la frecuencia.



Filtros

Es un circuito electrónico que permite pasar ciertas frecuencias mientras filtra otras frecuencias. Por lo tanto, un filtro puede eliminar las frecuencias deseadas de diferentes señales que también incluyen frecuencias irrelevantes o no deseadas. Es un elemento que discrimina una determinada frecuencia o gama de frecuencias de una señal eléctrica que pasa a través de él, pudiendo modificar tanto su amplitud como su fase.

Hay diferentes tipos de filtros disponibles como un filtro activo, filtro de paso de banda, filtro pasivo, filtro de paso bajo, filtro adaptado, filtro de paso alto, etc. Cada filtro tiene su propia función y aplicación.

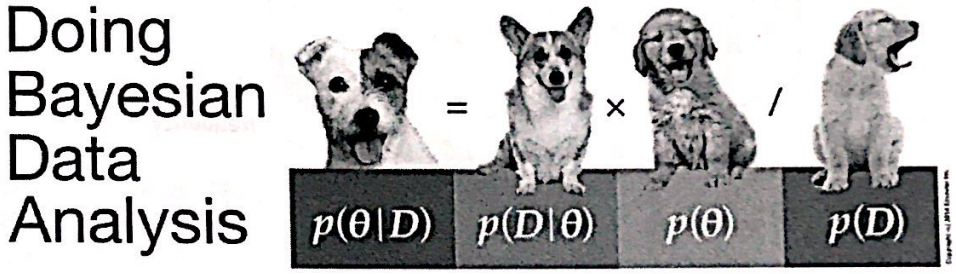
- Filtro de paso bajo (LPF): Se caracteriza por permitir el paso de las frecuencias más bajas y atenuar las frecuencias más altas. La entrada recoge todas las frecuencias, mientras que en la señal de salida solo se encontrarán las frecuencias que permita el filtro.
- Filtro de paso alto (HPF): Permite solo el paso de las frecuencias situadas por encima de una frecuencia particular, llamada frecuencia de corte y atenúa las frecuencias por debajo de esta.
- Filtro de paso banda (BPF): Es aquel que permite el paso de frecuencias comprendidas en un determinado rango de frecuencias, es decir, entre una frecuencia de corte superior (FH) y una inferior (FL).

BUSCAR MACHING FILTERS. (PAPERS).

Scanned with CamScanner

Modelo Bayesiano

La premisa básica del método bayesiano es que las declaraciones de probabilidad no se limitan a los datos, sino que se pueden hacer para los parámetros del modelo y los propios modelos. Las inferencias se realizan produciendo funciones de densidad de probabilidad (pdf); en particular, los parámetros del modelo se tratan como variables aleatorias.



Criterios de información para la selección de modelos

- Criterio de información de Akaike (AIC)

Es un estimador de la calidad relativa del modelo que tiene en cuenta su complejidad. AIC penaliza los modelos complejos en favor de los sencillos para evitar el sobreajuste.
Es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.
AIC maneja una solución de compromiso entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos.
AIC no proporciona una prueba de un modelo en el sentido de probar una hipótesis nula, es decir AIC no puede decir nada acerca de la calidad del modelo en un sentido absoluto. Si todos los modelos candidatos encajan mal, AIC no dará ningún aviso de ello. En el caso general, el AIC es:

$$AIC = 2k - 2\ln(L)$$

donde k es el número de parámetros en el modelo estadístico , y L es el máximo valor del Likelihood para el modelo estimado.
Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC. Por lo tanto AIC no solamente recompensa la bondad de ajuste, sino también incluye una penalidad, que es una función creciente del número de parámetros estimados. Esta penalización desalienta el sobreajuste.

Scanned with CamScanner

- Criterio de información Bayesiano (BIC)

Criterio de Información Bayesiano (BIC)es una medida de bondad de ajuste de un modelo estadístico, y es a menudo utilizado como un criterio para para la selección de modelos entre un conjunto finito de modelos. Se basa en la función de probabilidad logarítmica (LLF) y está estrechamente relacionado con el criterio de información de Akaike.
Al igual que en la AIC, BIC introduce un término de penalización para el número de parámetros en el modelo, pero la pena es mayor que uno en la AIC.
En general, el BIC es definido como:

$$\mathbf{BIC=k\times ln(n)-2\times ln(L)}$$

donde **k** es el número de parámetros del modelo y **ln(L)** es la función de log del Likelihood para el modelo estadístico.
Dados dos modelos estimados, el modelo con el menor valor de BIC es preferido; un BIC bajo implica un número menor de variables explicativas, mejor ajuste, o ambos.
Es importante tener en cuenta que el BIC se puede utilizar para comparar los modelos estimados sólo cuando los valores numéricos de la variable dependiente son idénticos para que todas las estimaciones sean comparadas.
BIC ha sido ampliamente utilizado para la identificación del modelo en series de tiempo y la regresión lineal. Puede, sin embargo, se aplica muy ampliamente a cualquier conjunto de modelos basados en una máxima probabilidad.

Función de correlación

En secciones anteriores, describimos la búsqueda de estructura dentro de los datos puntuales utilizando la estimación de densidad y la identificación de conglomerados. Para los procesos puntuales, una extensión popular de estas ideas es el uso de funciones de correlación para caracterizar hasta qué punto (y en qué escalas) la distribución de puntos difiere de una distribución aleatoria.

PREGUNTAR A MARIANO QUÉ MÁS DEBO SABER DE ESTO:

Scanned with CamScanner

¿Qué es matched filtered?

Quiero detectar algo. Filtro con una convolución
↓
con el complejo conjugado de sí mismo.

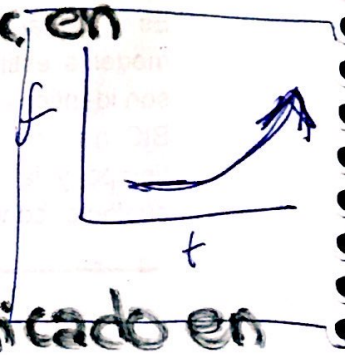
Paper → el filtro tiene info
de la forma y luminosidad
del cúmulo

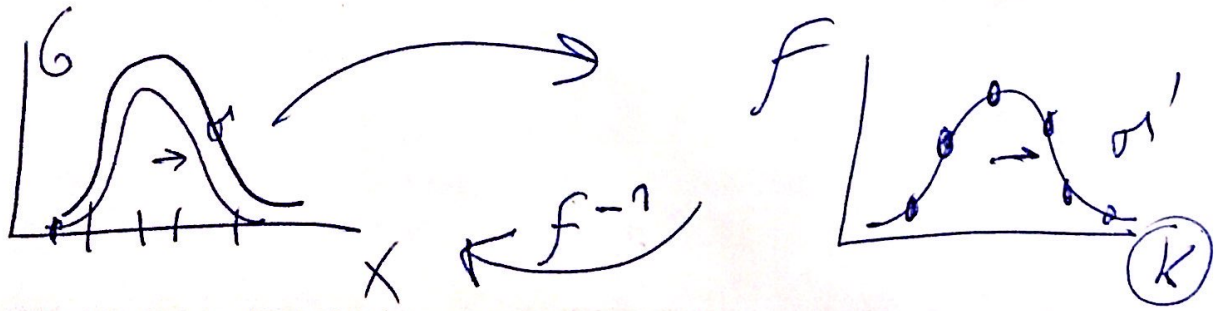
→ Wolter.

NUMERICAL → TRANSF.

"bonditas" ¿Que las distinguen de Fourier?

- ¿Porque es rapida la rápidez?
- ¿Cómo están ordenadas las frec. en Fourier?



- Nombres frec. altas y bajas.
- Teorema de convolución → aplicado en discreto y aplicado a la rápida.
- Estadísticas importante en el espacio de Fourier (transformada de correla-ción).

FFTW