(7)

$\ell_2(c) = \sum_{i=1}^{n} (x_i - c)^2$ QUIERO C TAL QUE $\ell_2(c)$ SEA MÍNIMA.

OSEA QUE ESTA SUMA SEA MÍNIMA.

¿QUÉ ES ESTA SUMA?

$(x_i - c)^2$ SERÍA COMO LA DISTANCIA DE LOS X_i A ESA CONSTANTE C AL CUADRADO

$\sum_{i=1}^{n} (x_i - c)^2$ ENTONCES ACÁ ESTOY SUMANDO LAS DISTANCIAS DE LAS V.A. X_i A LA CONSTANTE C.

ENTONCES LO QUE ESTOY BUSCANDO ES: CUAL ES LA CONSTANTE C QUE MEJOR APROXIMA A TODOS MIS DATOS A LA VEZ.

↳ NO ME SIRVE SI ESTÁ MUY CERCA DE UNA PERO MUY LEJOS DE LA OTRA. => POR QUE IGUAL LA SUMA ME QUEDARÍA GRANDE.

(TENGO 10 X_i)

ESTAMOS BUSCANDO CUAL ES LA CONSTANTE QUE MINIMIZA LO MEJOR POSIBLE PARA TODAS LAS X_i EL CUADRADO DE LAS DISTANCIAS DE LAS X_i A C

REGRESIÓN LINEAL: SI TUVIERA UN DATO MÁS, POR EJ: X_{11}

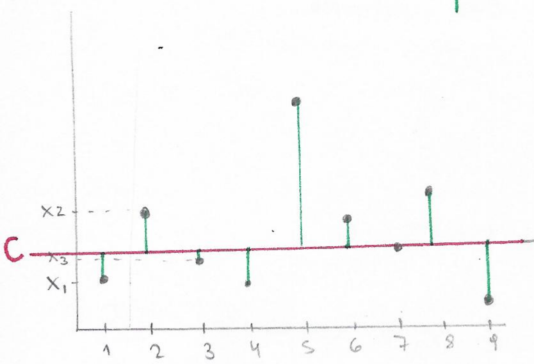
SI EFECTIVAMENTE ENCONTRÉ LA CONSTANTE C QUE MINIMIZA LA DISTANCIA DE TODOS LOS X_i, SE PODRÍA PENSAR QUE SI VIENE UN DATO NUEVO (X_{11}) NO VA A ESTAR TAN LEJOS DE LA CONSTANTE C.

SI TENGO QUE INTENTAR ENCONTRAR LA CONSTANTE QUE MEJOR APROXIMA A LOS VALORES DE UNA V.A. ESO EN GENERAL SE PARECERÁ A LA ESPERANZA (DE ESAS V.A.)

POR EJ: SI TUVIERA UNA MONEDA CARGADA, QUE SALE CARA CON PROBA. 0,2 Y TIRO LA MONEDA 600.000 VECES, Y QUIERO SABER CUANTAS SALIERON CARAS, NO VOY A INTENTAR ADIVINAR EL NÚMERO EXACTO, VOY A HACER (0,2).600000 Y DECIR QUE EL RESULTADO ANDARÁ POR AHÍ

↳ ESO TIENE QUE VER CON USAR LA ESPERANZA COMO LA CONSTANTE QUE MEJOR APROXIMA A LA FUNCIÓN.

DISTANCIAS



→ ESTAMOS BUSCANDO LA CONSTANTE C QUE MINIMIZA LAS DISTANCIAS SUMADAS

LA FUNCIÓN BUSCADA $\ell_2(c)$ ES LA QUE VA A SUMAR CADA LARGO DE LAS LÍNEAS VERDES.

⇒ LO QUE QUIERO ES:

CUAL ES EL C A CUYA ALTURA ME CONVIENE PONER LA RECTA PARA QUE LAS DISTANCIAS SEAN LA MENOR POSIBLE (AL SUMARLAS) (DE TODAS A LA VEZ).

TÍPICAMENTE SERÁ ALGO EQUILIBRADO EN EL MEDIO DE TODAS LAS X_i

SI QUIERO MINIMIZAR LA DISTANCIA AL X_1 PONGO LA CONSTANTE A ESA ALTURA Y YA ESTÁ.
PERO LO QUIERO ES MINIMIZAR TODAS LAS DISTANCIAS A LA VEZ.
ENTONCES SI PONDRÍA A C EN X_1, SE ME ALEJARÍA MUCHO DE X_5 POR EJEMPLO

INTUITIVAMENTE, EL NÚMERO QUE MEJOR APROXIMA A TODOS A LA VEZ ES EL PROMEDIO.
(QUE CASUALMENTE SE PARECE A LA ESPERANZA SI EL n ES GRANDE (LGN))

LO DEMOSTRAMOS:

Minimicemos $\ell_2(c)$:

→ SU ÚNICO EXTREMO POSIBLE ES SU MÍNIMO
CUANDO DERIVE E IGUALE A 0 NO APARECERÁ MÁXIMOS.
SI YO AGARRO UN C MUY ARRIBA, LA DISTANCIA PUEDE SER TAN GRANDE COMO YO QUIERA

OSEA BUSQUEMOS EL MÍNIMO DE $\sum_{i=1}^{n}(x_i - c)^2$

⇒ SI HAY UN PUNTO EXTREMO ES UN MÍNIMO (PQ LA DISTANCIA SIEMPRE PUEDE SER MAYOR, TOMANDO C ARBITRARIAMENTE GRANDE)

DERIVO E IGUALO A 0:

[C ES LA VARIABLE DE MI FUNCIÓN] LAS X_i SON COMO NÚMEROS CONSTANTES

SERÍA TIPO COMO $(4 - x^2) + (5 - x^2) + \dots$

$$\ell'(c) = \sum_{i=1}^{n} 2.(-1)(x_i - c) = 0$$

$$\iff \sum_{i=1}^{n}(x_i - c) = 0 \iff \left(\sum_{i=1}^{n} x_i\right) - n.c = 0$$

$(x_1 - c) + (x_2 - c) + \dots (x_n - c)$

$\left(\sum_{i=1}^{n} x_i\right) - nc$

$$\iff c = \frac{\sum_{i=1}^{n} x_i}{n}$$

Obs: ACÁ NO HAY NADA ALEATORIO.
UNA VEZ QUE ME DAN LOS X_i QUE SÍ SON ALEATORIOS, TODO EL PROCEDIMIENTO PARA HALLAR C FUE DETERMINÍSTICO.
OSEA NO DEPENDE DE LOS DATOS NI DE NADA.

SI QUIERO MINIMIZAR ESA DISTANCIA, MATEMÁTICAMENTE PROBÉ QUE ME DEBERÍA AGARRAR EL PROMEDIO.

EL MEJOR ESTIMADOR LINEAL CONSTANTE PARA EL CONJUNTO DE DATOS ES CONSIDERAR EL PROMEDIO.

$f_1(c) = \sum_{i=1}^{n} |x_i - c|$ MIDE MÁS POSTA CUAL ES LA DISTANCIA, PORQUE NO LA MODIFICA

ME DICE LA DISTANCIA Y PUNTO.

PERO <u>NO</u> ES DERIVABLE.

EN CAMBIO

$f_2(c) = \sum_{i=1}^{n} (x_i - c)^2$ DISTORSIONA LAS DISTANCIAS, PERO <u>SÍ</u> ES DERIVABLE.