Regresión lineal

Familia: métodos geométricos (lineales)

Tarea: regresión

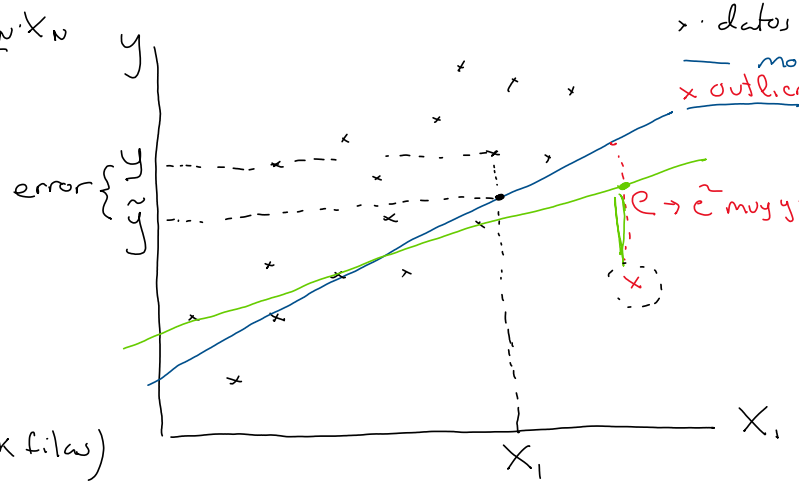
Idea matemática: escoger los coeficientes (A y B_N) tales que minimizan el error cuadrático medio (dada una relación lineal entre atributo y target)

Algoritmo: $Y = A + B_1 \cdot X_1 + B_2 \cdot X_2 + \dots + B_N \cdot X_N$

→ mucho sesgo: presupone una relación lineal y constante

↳ genera underfitting

↳ utilizamos si tenemos pocos datos (10K - 20K filas)



$error = y - \tilde{y}$

$\min \sum_{i=1}^{n} e_i^2$

↳ penaliza mucho errores grandes

↳ es sensible a outliers

→

Regresión logística

Familia: same

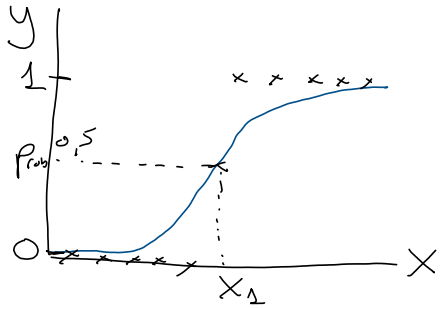
Tarea: clasificación

Idea: same

diferente:

$\tilde{y}$ debe ser entre 0 y 1

↳ utilizamos una sigmoide: $\tilde{y} = \frac{1}{1 + e^{-(A+Bx)}}$



Data Understanding: miraremos correlación entre variables

Correlación entre categóricas: test chi2

Ejemplo

id	cat1	cat2
1	A	C
2	A	C
3	A	C
4	A	C
5	B	C
6	B	D
7	B	D
8	B	D
9	B	D
10	B	D

a simple vist parece que si cat1 = A, cat2 = C } mucha correlación
si cat1 = B, cat2 = D }

↓

cat1 \ cat2	C	D
A	4	0
B	1	5

H_0: no hay relación
H_1: hay relación
test mide la probabilidad de estas en concreto dada la H_0 → p-value ≤ 0.05 rechazamos H_0

En H_0, cat 1 y cat 2 son independientes
Calculamos algunos conteos:
cat 2 → 50% (C)
↘ 50% (D)
si son independientes, deberíamos tener que cuando cat 1 = A tengo la mitad de "C's y la mitad "D"s

$P(A) = 40\%$
$P(C) = 50\%$
$P(A, C) = P(A) \cdot P(C) = 20\%$
40% 50%
Realidad: $P(A, C) = 40\%$ ↙

poco probable que cat 1 y cat 2 sean independientes