Clustering languages

TABLE 12.3 NUMERALS IN 11 LANGUAGES

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
sev...	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Concordance of first letters= ...

	En	No	Da	Du	Ge	Fr	Sp	It	Pol	Hu	Fi
En	10	8	8	3	4	4	4	4	3	1	1
No	8	10	9	5	6	4	4	4	3	2	1
Da	8	9	10	4	5	4	5	5	4	2	1
Du	3	5	4	10	5	1	1	1	0	2	1
Ge	4	6	5	5	10	3	3	3	2	1	1
Fr	4	4	4	1	3	10	8	9	5	0	1
Sp	4	4	4	1	3	8	10	9	7	0	1
It	4	4	5	1	3	9	9	10	6	0	1
Pol	3	3	4	0	2	5	7	6	10	0	1
Hu	1	2	2	2	1	0	0	0	0	10	2
Hu	1	1	1	1	1	1	1	1	1	2	10

Clustering languages

TABLE 12.3 NUMERALS IN 11 LANGUAGES

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
sev...	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Concordance of first letters= ...

	En	No	Da	Du	Ge	Fr	Sp	It	Pol	Hu	Fi
En	10	8	8	3	4	4	4	4	3	1	1
No	8	10	9	5	6	4	4	4	3	2	1
Da	8	9	10	4	5	4	5	5	4	2	1
Du	3	5	4	10	5	1	1	1	0	2	1
Ge	4	6	5	5	10	3	3	3	2	1	1
Fr	4	4	4	1	3	10	8	9	5	0	1
Sp	4	4	4	1	3	8	10	9	7	0	1
It	4	4	5	1	3	9	9	10	6	0	1
Pol	3	3	4	0	2	5	7	6	10	0	1
Hu	1	2	2	2	1	0	0	0	0	10	2
Hu	1	1	1	1	1	1	1	1	1	2	10

$$\langle n_H(t+1)\rangle \geq \overbrace{\frac{\bar{f}_H(t)}{\bar{f}(t)}\, n_H(t)}^{S_s(H)} \overbrace{\left(1 - P_c \frac{\delta(H)}{l-1}\right)}^{S_c(H)} \overbrace{(1 - P_m)^{O(H)}}^{S_m(H)}$$

$\langle n_H(t+1)\rangle$: # ~~de~~ esperado de instancias de H.

$\bar{f}_H(t)$: Fitness medio del esquema

$\bar{f}(t)$: Media del fitness.

$n_H(t)$: # de cromosomas que implementa el esquema.

P_c: Prob. de cruce = 1

$\delta(H)$: long. del esquema.

l: long. de los cromosomas

P_m: Prob. de mutar

$O(H)$: # de bits que vienen definidos (orden del esquema)

✱ Cuando los bits están en los extremos $S_c = \emptyset$

Ss : Prob. de selección

Sc : Prob. de cruce
que sobreviva en el

SM : Prob. de mutación.
que sobreviva a la

$\delta(H)$: long. del esquema. (long. de máx. de un bit definido a otro).

Ejm :

$$\left.\begin{array}{l} 1 * * * * = \emptyset \\ * 1 0 * * = 1 \\ 1 * * * 0 = 4 \end{array}\right\} l = 5$$

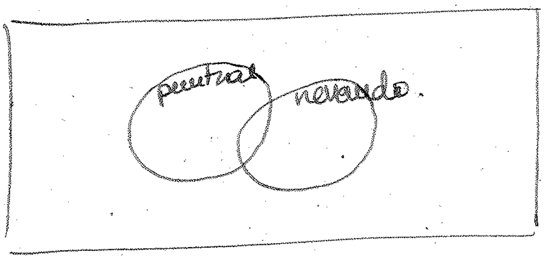
e l a p d

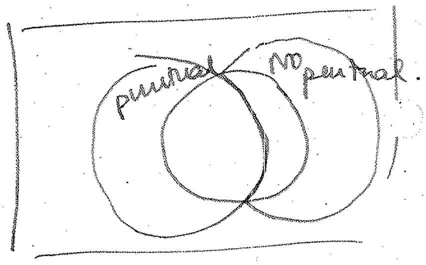
AGE	PRESCRIPTION	ASTIGMATISM	TEAR_RATE	LENSES
young 0	myope 1	no 2	reduced 3	no 4
young	myope	no	normal 5	soft 6
young	myope	yes	reduced	no
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	no
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	no
young	hypermetrope	yes	normal	hard
pre-pres	myope	no	reduced	no
pre-pres	myope	no	normal	soft
pre-pres	myope	yes	reduced	no
pre-pres	myope	yes	normal	hard
pre-pres	hypermetrope	no	reduced	no
pre-pres	hypermetrope	no	normal	soft
pre-pres	hypermetrope	yes	reduced	no
pre-pres	hypermetrope	yes	normal	no
presbyopic	myope	no	reduced	no
presbyopic	myope	no	normal	no
presbyopic	myope	yes	reduced	no
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	no
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	no
presbyopic	hypermetrope	yes	normal	no

$$P(\text{cond. Adv}) = 98\% \times 20\% + 2\% \cdot 70\% = $$
$$0'196 + 0'014 = 0'21 = 21\%$$

$$P(\overset{A}{\text{Puntual}} \mid \overset{B}{\text{Nieve}}, I) = \frac{P(\text{Nieve} \mid \text{Puntual}, I) \cdot P(\text{Puntual})}{P(\text{nieve})} =$$

$$= \frac{20\% \cdot 98\%}{\cancel{0}21\%} = \frac{0'196}{0'21} = 0'93\cancel{3} \; 93\%$$

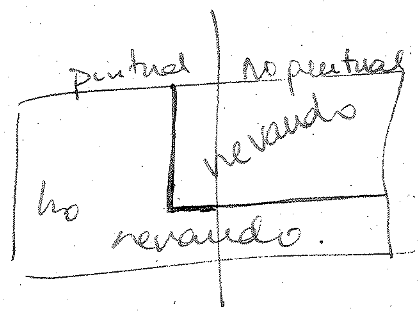


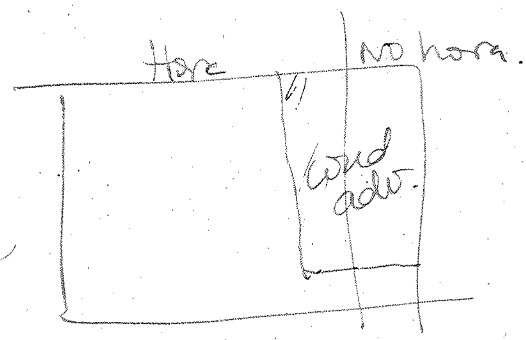


$A \rightarrow B.$

$\neg B \rightarrow \neg A.$

$$P(B \mid A, A \rightarrow B) = \frac{P(A \mid B, A \rightarrow B) \cdot P(A)}{P(B)}$$





(H) D,

parcial 5

REGRESIÓN LOGÍSTICA

$\bar{w} \rightarrow$ vector aleatorio entre $[-0'5, 0'5]$ (lo pueden dar)

$\bar{x} \rightarrow$ valores de las columnas donde $\boxed{x_0 = 1}$

x_1	x_2	x_3	x_4	t
...	...	...	...	...

sigmoidal : $\sigma(\bar{w}^T \bar{x}_i) = - \dfrac{1}{1 + e^{\bar{w}^T \bar{x}_i}}$

$\hookrightarrow \bar{w} = (w_0, w_1, w_2)$

$\bar{w}^T = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$

AJUSTE DE LA RECTA $\bar{w}$: $\bar{w} = \bar{w} - \eta (\sigma_i - t_i) \bar{x}_i$

$\rightarrow$ sigmoidal

vector que define el hiperplano

$\hookrightarrow$ cte entrenamiento

$\hookrightarrow$ clase

(*) ÉPOCAS : permite ajustar los valores η y se pueden aleatorizar los datos por cada época para mejorar. También se puede variar w_0 para que la recta no crezca demasiado.

EJEMPLO :

$\bar{w} = (0'5, 0'6, -0'2)$ $\quad \bar{w}^T = \begin{pmatrix} 0'5 \\ 0'6 \\ -0'2 \end{pmatrix}$

$\bar{x}_i = (1, \overset{x}{3}, 1)$ $\quad \eta = 2$

$\bar{w}^T \bar{x}_i = \begin{pmatrix} 0'5 \\ 0'6 \\ -0'2 \end{pmatrix} \cdot (1, 3, 1) = 0'5 \cdot 1 + 0'6 \cdot 3 + (-0'2) \cdot 1 = 2'1$

$\sigma(2'1) = - \dfrac{1}{1 + e^{2'1}} = \underline{\underline{0'10}}$

$\begin{cases} \text{clase } 1 = 1 \\ \text{clase } 2 = 0 \end{cases}$

ejemplo de clase 1 $\downarrow$

Actualizamos $\bar{w}$: $\bar{w} = (0'5, 0'6, -0'2) - 2(0'1 - 1) \cdot (1, 3, 1) =$

$= (0'5, 0'6, -0'2) - (\underset{1'8}{\cancel{\;}}, \underset{5'4}{\cancel{\;}}, 1'8) = (2'3, 6, 1'6)$

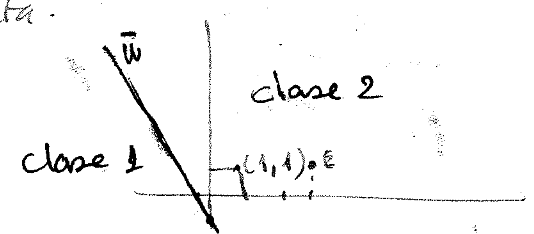
calculamos la recta : $w_0 + w_1 x + w_2 y \rightarrow 2'3 + 6x + 1'6y$

$x = 0 \rightarrow y = -1'43$

$y = 0 \rightarrow x = -0'38$

calculamos $(0,0)$ $\quad \sigma(2'3) = 0'09$

$0'09 < 0'5 \rightarrow C_2$



Prop. por MAP.

Cambia la actualización de $\bar{w}$

$$\bar{w} = \bar{w} - \eta\,(\bar{o}_i - t_i)\bar{x}_i - \frac{\eta}{N\sigma^2}\,\bar{w}$$

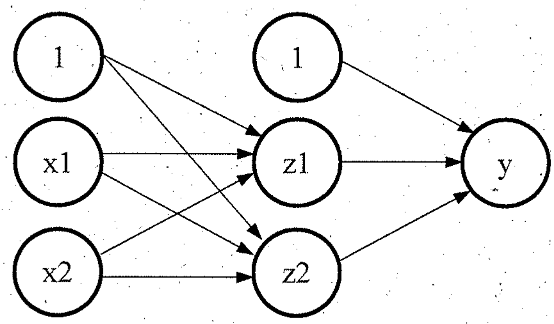
→ varianza (la dan)

→ nº ejemplos

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

1. Given the following instances, obtain a valid set of weights for the neural network architecture shown in figure in order to get all examples correctly classified and compute the values of the z variables. The activation functions are all step functions.

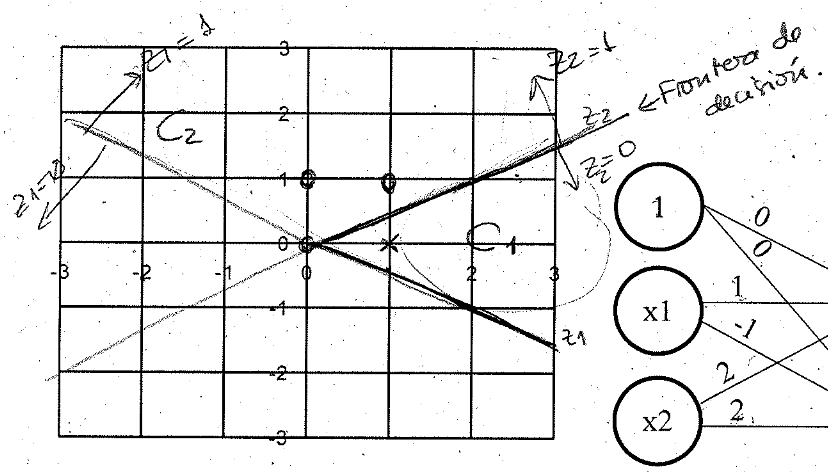
x_1	x_2	t	z_1	z_2
0	0	0		
0	1	0		
1	0	0		
1	1	0		
0,5	0,5	1		

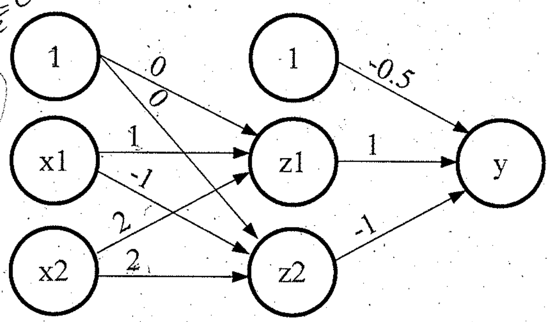


2. Draw the decision boundary of the following neural network

$z_1 = \sigma(x_1 + 2x_2)$

$z_2 = \sigma(-x_1 + 2x_2)$





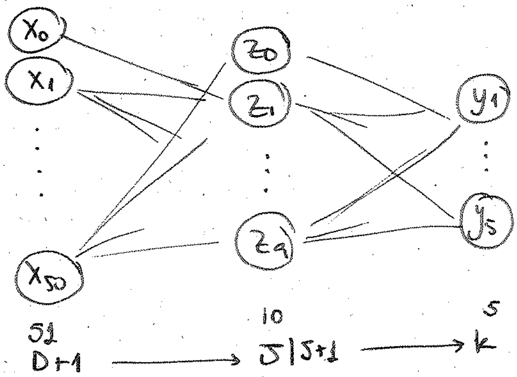
z_1	z_2	$y = \sigma(-0'5 + z_1 - z_2)$	clase
0	0	0	C_2
0	1	0	C_2
1	0	1	C_1
1	1	0	C_2

3. We would like to solve a classification problem with 5 classes, 50 attributes and 1000000 training instances using a neural network with one hidden layer containing 9 neurons (z_1, z_2,...,z_9). Ten training epochs will be executed. Compute the number of weight updates that have to be carried out during the traiing phase.

clasificar: $O(DJ + Jk)$
entrenamiento:

nº pesos $= (D+1) \times J + (J+1) \times k$
nº epocas $= 10$ $\quad N = 100.000$
no. actualizaciones = epocas × N × no.pesos

4. Compute the number of necesary operations to classify an instance using the network described in the previous exercice.



productos: $(D+1) \times J$ ⋮ $(J+1) \times k$
sumas: $D \times J$ ⋮ $J \times k$
+ J sigmoidales
+ k softmax

orden ejecución
$O(DJ + Jk)$

$51 \times 9 + 10 \times 5 = 509$ productos
$50 \times 9 + 9 \times 5 = 495$ sumas

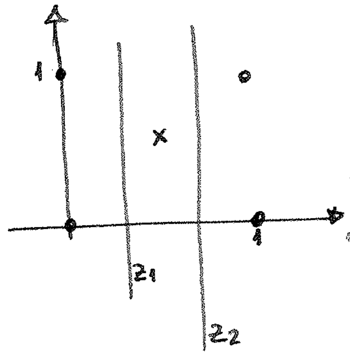
5. We would like to train a Rosenblatt perceptron for a problem with 4 attributes. The weights are w={0.1 -0.1 0 0.1 -0.1}. Compute the weight update for the example x={1 1 0 1} of class t=1 using 0.25 as learning coefficient.

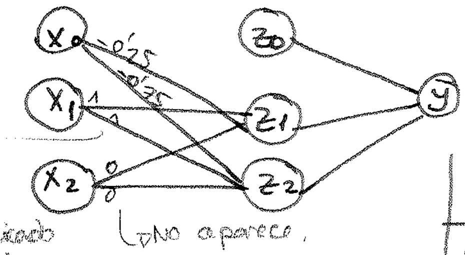
$$y(\bar{x}) = \theta\left(\sum_{i=0}^{D} x_i w_i\right) = \theta(-0'1) = 0$$

$$\sum_{i=0}^{D} x_i w_i = 0'1 - 0'1 + 0 + 0 - 0'1 = -0'1$$

$$\bar{w} = \bar{w} - \eta\,(y(\bar{x}) - t)\,\bar{x} = \begin{pmatrix} 0'1 \\ 0'1 \\ 0 \\ 0'1 \\ -0'1 \end{pmatrix} - 0'25\,(0-1)\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0'35 \\ 0'15 \\ 0'25 \\ 0'1 \\ 0'15 \end{pmatrix}$$

Ej 1 Hoja de ejercicios





x_1 no está multiplicado por nada.

↳ No aparece.

Fn activación $= \theta$

$x_1 = 0'25$

$$\left.\begin{aligned} z_1 &= \theta(x_1 - 0'25) \\ z_2 &= \theta(x_1 - 0'75) \end{aligned}\right\}$$

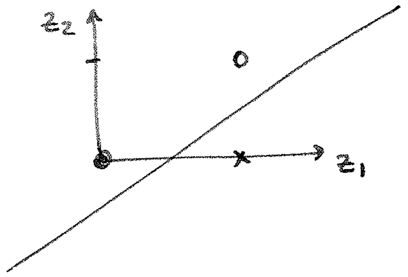
clasificación ↑ variables

x_1	x_2	t	z_1	z_2
0	0	0	0	0
0	1	0	0	0
1	0	0	1	1
1	1	0	1	1
0'5	0'5	1	1	0

Sustituimos x_1 en la ec.

Si es $t = 1$ y $- = 0$

→ Nueva clasificación a partir de la tabla.



$z_1 = 0 \rightarrow z_2 = -0'5$

$z_2 = 0 \rightarrow z_1 = 0'5$

recta: $z_1 - z_2 - 0'5 = 0$

$-z_1 + z_2 + 0'5 = 0$

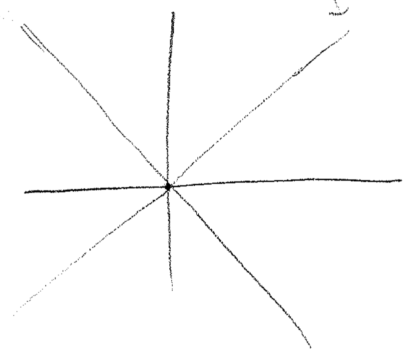
} Elegimos la que coincide con la salida t.

coincide con t.

$$\theta(z_1 - z_2 - 0'5).$$

Ejercicio 2

hacer los br...



$z_1 = \theta(x_1 + 2x_2)$.

$z_2 = \theta(-x_1 + 2x_2)$

$x_1 + 2x_2 = 0$; $x_2 = -\frac{1}{2}x_1$

La vie au quotidien

1 LEXIQUE

Complétez cet article à l'aide de mots choisis dans la liste suivante. Faites les transformations nécessaires.

look – porter – soigné – vêtement – tenue vestimentaire – classique – présentation – apparence – image – négligé – pied – confort – adapté – excentrique – chic – code vestimentaire – s'habiller – fonction

Quand l'habit fait le moine...

« Dis-moi ce que tu .... et je te dirai qui tu es. » Loin de n'être que secondaire, la .... a une grande importance lors d'un entretien de recrutement. Elle fait partie des signes perçus par votre interlocuteur pour se faire une bonne ou une mauvaise .... de vous.

Rien de plus superficiel que le ...., pensez-vous ? Erreur ! Votre .... en dit parfois plus sur vous qu'un long discours. Lors d'un entretien d'embauche, votre .... joue donc un vrai rôle. Si certaines entreprises en parlent dans leurs annonces avec des formules du type « présentation .... exigée », d'autres ne l'affichent pas forcément, mais y accordent tout autant d'importance.

La tête de l'emploi

Un look .... à l'entreprise vous permettra d'abord de vous rassurer et en même temps d'indiquer à votre interlocuteur que vous vous fondez dans la culture de l'entreprise.

Prenez également en considération votre .... Cela peut paraître une évidence, mais on ne .... pas de la même façon quand on est banquier, informaticien ou créatif.

Si vous ignorez les .... en usage dans l'entreprise où vous vous rendez, restez ...., évitez les tenues ....

Deux excès sont à éviter. Trop ...., vous prenez le risque que votre interlocuteur se sente dévalorisé. Trop ...., vous pourriez le choquer.

Il est aussi essentiel que vous soyez bien dans vos baskets... ou dans votre tailleur-pantalon ! Le .... est un élément très important dont il faut tenir compte : ne portez pas de .... qui vous gêneraient ou des chaussures auxquelles vous n'êtes pas habitué. Vous risqueriez de vous retrouver dans des situations embarrassantes et de vous concentrer sur vos .... qui vous font souffrir plutôt que sur les questions du recruteur !

D'après Rebondir n° 127, janvier 2006.

1 **(TAXIS)** Un taxi golpea a una persona de noche y huye. En la ciudad operan dos compañías de taxis: la verde y la azul. El 85 % de los taxis de la ciudad son verdes y el 15 % restante, azules.

1. ¿Cuál es la probabilidad de que el taxi del accidente fuera azul?
2. Aparece una testigo que identifica el taxi como azul. El jurado estima la fiabilidad de la testigo en un 80 %. ¿Cuál es la probabilidad ahora de que el taxi del accidente fuera azul?
3. Decisión final usando el criterio de máxima verosimilitud y el citerio de máxima probabilidad a posteriori.

2 **(NO CAUSAL)** Las probabilidades pueden representar conexiones lógicas, no causales. Consideremos una urna con 6 bolas blancas y 6 negras.

1. Si en la primera extracción se ha eliminado una bola blanca, ¿Cuál es la probabilidad de extraer una bola blanca en una segunda extracción?
2. Por el contrario, supongamos que la segunda extracción ha sido blanca, ¿Cuál es la probabilidad de que la primera fuera blanca también?

3 **(LLUVIA)** Según el hombre del tiempo la probabilidad de lluvia hoy es del 20 %. Estamos en un sótano sin ventanas y no podemos saber que tiempo hace fuera. Sin embargo, vemos entrar a alguien llevando un parágua. Sabiendo que la probabilidad de que alguien lleve un paraáguas y esté lloviendo es del 70 % y solo del 10 % en caso contrario ¿Cuál es la probabilidad de que esté lloviendo?

4 **(BOLSAS)** Hay dos bolsas: A y B. La bolsa A contiene 2 bolas negras y 3 blancas. La B, 3 negras y 2 blancas. Se selecciona una bolsa al azar y se extrae una bola.

1. Calcular la probabilidad de que la bola sea negra.
2. Si la bola obtenida ha sido negra calcular qué bolsa es la más probable que hayamos elegido.
3. Decisión final usando el criterio de máxima verosimilitud y el citerio de máxima probabilidad a posteriori

5 **(BOLSAS 2)** Hay dos bolsas: A y B. La bolsa A contiene 2 bolas negras y 3 blancas. La B, 3 negras y 2 blancas. Se selecciona una bolsa al azar, teniendo en cuenta que la bolsa A tiene un 75 % de probabilidad de ser elegida, y se extrae una bola.

1. Calcular la probabilidad de que la bola sea negra.
2. Si la bola obtenida ha sido negra calcular qué bolsa es la más probable que hayamos elegido.
3. Decisión final usando el criterio de máxima verosimilitud y el citerio de máxima probabilidad a posteriori

Verosimilitud

$$p(a=1.5\,|\,i=set)=0$$

$$p(a=1.5\,|\,i=ver)=\frac{17}{50}$$

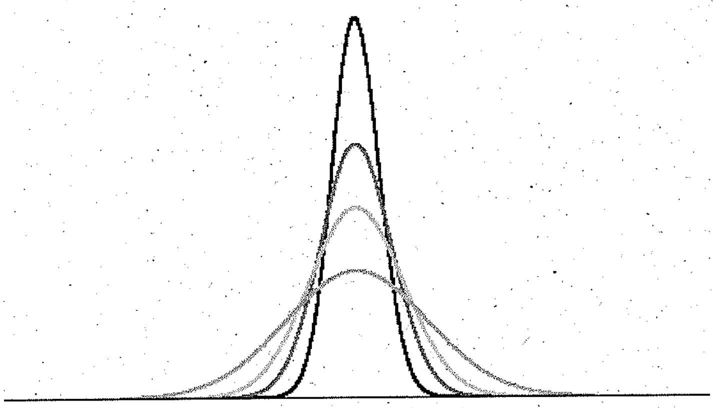
$$p(a=1.5\,|\,i=set)=\frac{3}{50}$$

Anchura de pétalos	setosa	versicolor	virginica
0 - 0.2	1/10		
0.2 - 0.4	18/25		
0.4 - 0.6	4/25		
0.6 - 0.8	1/50		
0.8 – 1.0			
1.0 – 1.2		1/5	
1.2 – 1.4		9/25	
1.4 – 1.6		17/50	3/50
1.6 – 1.8		2/25	1/25
1.8 – 2.0		1/50	8/25
2.0 – 2.2			6/25
2.2 - 2.4			11/50
2.4 – 2.6			3/25
2.6 – 2.8			
2.8 – 3.0			

Distribución normal

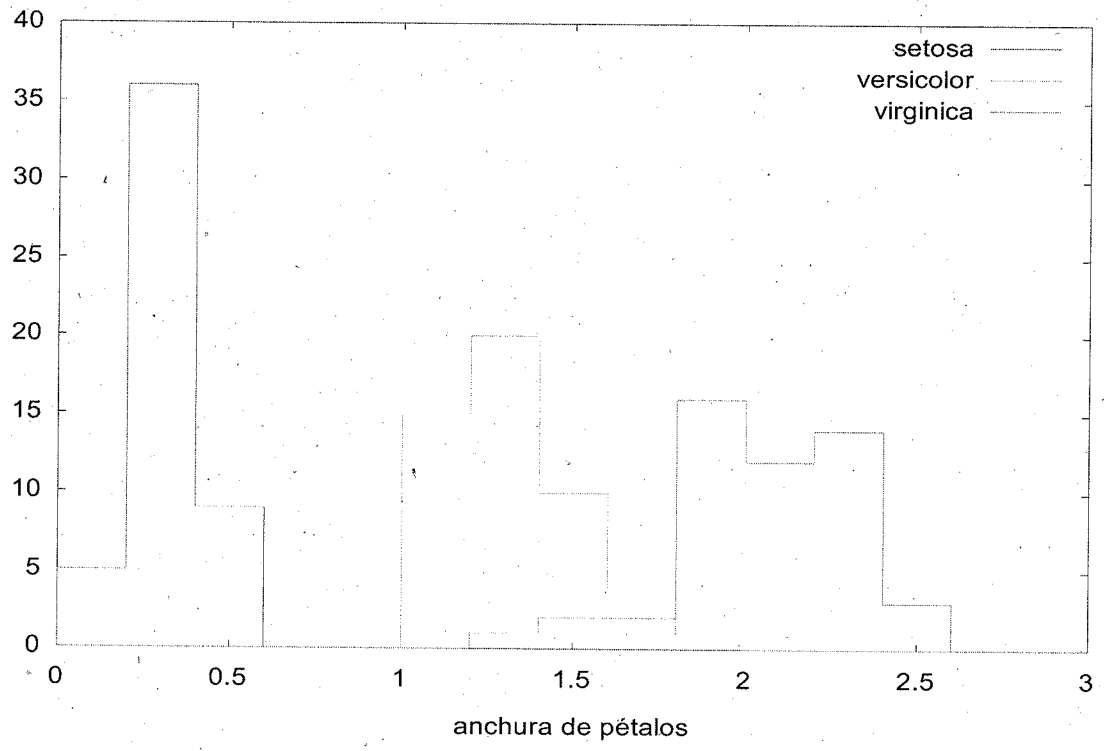
$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-m)^2}{2\sigma^2}}$$

(a) Distribuciones normales con distinta desviación estándar e igual media



Naive Bayes para atributos cuantitativos

¿A qué clase debemos asignar una flor con anchura 1.5?



Discretización de la anchura de pétalos

Anchura de pétalos	setosa	versicolor	virginica
[0 - 0.2)	5		
0.2 - 0.4	36		
0.4 - 0.6	8		
0.6 - 0.8	1		
0.8 – 1.0			
1.0 – 1.2		10	
1.2 – 1.4		18	
1.4 – 1.6		17	3
1.6 – 1.8		4	2
1.8 – 2.0		1	16
2.0 – 2.2			12
2.2 - 2.4			11
2.4 – 2.6			6
2.6 – 2.8			
2.8 – 3.0			

12 sept '13

— PROBABILIDAD. Introducción

• Lógica deductiva

* Silogismos fuertes

	A	B	$A \to B$
FUERTE	0	0	1
	0	1	1
	1	0	0
FUERTE	1	1	1

PREMISA: Si $A \to B$ y observamos A, entonces B es cierto.
Además, si observamos $\neg B$ entonces A no es cierto.

* Silogismos débiles

Si B es cierto entonces A es <u>más probable</u>

Si A es falso entonces B es <u>menos probab</u>

• Conceptos de probabilidad

DOMINIO: Todos los posibles resultados de un experimento aleatorio.

VARIABLE ALEATORIA: Variable que identifica al experimento aleatorio

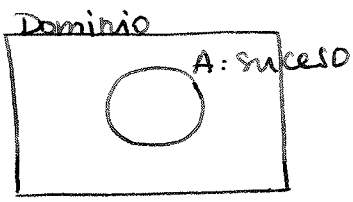
SUCESO: valor que toma la variable aleatoria al realizar el experimento aleatorio.

ej: TIRAR UN DADO

Dominio = $\{1, 2, 3, 4, 5, 6\}$

Variable = X

Sucesos = $X = 1, \ldots, X = 6$

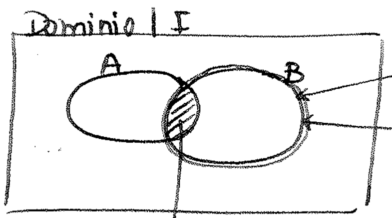


$P(A|I)$ donde A es el suceso e I la información

· **REGLAS**

1) $P(A|I) \in [\emptyset, 1]$
2) $P(\text{true}) = 1$, $P(\text{false}) = \emptyset$
3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4) Regla de la suma: $P(A|I) + P(\bar{A}|I) = 1$
5) Regla del producto: $P(A,B|I) = P(A|B,I) \cdot P(B|I)$

$\hookrightarrow P(A|B,I) = \dfrac{P(A,B|I)}{P(B|I)}$



~~· PROBABILIDAD Y LÓGICA DEDUCTIVA~~ 13 sept '13

~~$P(A|I) + P(\bar{A}|I) = 1$ en el límite $P(A|I$~~

· Silogismo fuerte $I = A \Rightarrow B$

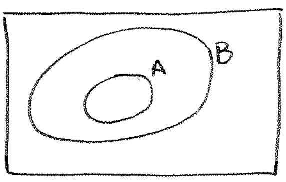
$$P(B|A, A \Rightarrow B) = \frac{P(A,B|A \Rightarrow B)}{P(A|A \Rightarrow B)} = \frac{P(A|A \Rightarrow B)}{P(A|A \Rightarrow B)} = 1$$

↑ aplicando la regla del producto.

porque si se cumple A, en $A \Rightarrow B$, B se va a cumplir

~~P(B~~

$$P(A|\bar{B}, A \Rightarrow B) = \frac{P(A,\bar{B}|A \Rightarrow B)}{P(\bar{B}|A \Rightarrow B)} = \frac{\emptyset}{P(\bar{B}|A \Rightarrow B)} = \emptyset$$



- <u>Teorema de Bayes</u>

$$P(A \mid B, I) = \frac{P(B \mid A, I)\, P(A \mid I)}{P(B \mid I)}$$

aplicándolo a patrones, B = Datos (D), A = Hipótesis (H

$$P(H \mid D, I) = \frac{P(D \mid H, I)\, P(H \mid I)}{P(D \mid I)}$$

$P(H \mid I) \equiv$ Probabilidad <u>a priori</u>, confianza en H antes de los datos

$P(D \mid H, I) \equiv$ <u>Verosimilitud</u> de los datos D si H es cierta.

$P(H \mid D, I) \equiv$ Probabilidad <u>a posteriori</u>.

$P(D \mid I) \equiv$ Evidencia.

Problema de los Taxis

H = azul o H = verde

Un taxi golpea a una persona de noche y huye. En la ciudad operan 2 compañías de taxis: la verde y la azul. El 85% de los taxis son verdes, y el 15% azules. ¿Cuál es la prob. de que el taxi del accidente fuera azul? Aparece un testigo que identifica el taxi como azul. La fiabilidad del testigo es del 80%. ¿Cuál es la prob. ahora de que el taxi fuera azul? ¿Decisión final utilizando MAP y MV?

$P(H = \text{azul} \mid I) = 15\%$ $\quad P(H = \text{verde} \mid I) = 85\%$

D = testigo dice azul confianza del 80%

$$P(H = \text{azul} \mid D = \text{azul}, I) = \frac{\overbrace{P(D = \text{azul} \mid H = \text{azul}, I)}^{80\%}\, \overbrace{P(H = \text{azul} \mid I)}^{15\%}}{P(D = \text{azul} \mid I)} = 0'41$$

$$P(H = \text{verde} \mid D = \text{azul}, I) = \frac{\overbrace{P(D = \text{azul} \mid H = \text{verde}, I)}^{20\%}\, \overbrace{P(H = \text{verde} \mid I)}^{85\%}}{P(D = \text{azul} \mid I)} = \underline{\underline{0'59}}$$

Más probabl[e] que sea verde.

$P(D = \text{azul} \mid I) = P(D = \text{azul}, H = \text{azul} \mid I) + (D = \text{azul}, H = \text{verde} \mid I) =$

$= P(D = \text{azul} \mid H = \text{azul}, I)\, P(H = \text{azul} \mid I) + P(D = \text{azul} \mid H = \text{verde}, I)\, P(H = \text{verde} \mid I) =$

$= 0.2 \cdot 0.85 + 0.8 \cdot 0'15 = 0'29$

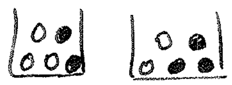
MAP: $\text{máx}\,(0'41, 0'59) = 0'59$ más prob. que fuera <u>verde</u>

MV: $\text{máx}\,(0'8, 0'2) = 0'8$ más prob. que fuera <u>azul</u>

18 Sept

Ej4 : Hay dos bolsas : A y B. La bolsa A contiene dos bolas negras y 3 blancas. La B, 3 negras y 2 blancas.

. Si la bola obtenida ha sido negra, calcular qué bolsa es la más probable que hayamos elegido.



A B

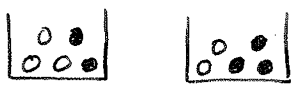
$$P(H=A \mid D=\bullet, I) = \frac{2/5 \cdot 1/2}{1/2} = 2/5$$

$$P(H=B \mid D=\bullet, I) = \underline{\underline{3/5}}$$

El resto.

más probable

Ej5: Hay dos bolsas: A y B



A B

75% de probabilidad

25% de prob.

· Calcular la prob. de que la bola sea negra:

$$P(D=\bullet \mid I) = 0'75 \cdot 2/5 + 0'25 \cdot 3/5 = 9/20$$

. Si la bola obtenida ha sido negra, calcular qué bolsa es la más probable que hayamos elegido.

$$P(H=A \mid D=\bullet, I) = \frac{P(D=\bullet \mid H=A, I)\, P(H=A \mid I)}{P(D=\bullet \mid I)} = \frac{2/5 \cdot 0'75}{9/20} = \underline{\underline{2/3}}$$

más prob.

$$P(H=B \mid D=\bullet, I) = \frac{P(D=\bullet \mid H=B, I)\, P(H=B \mid I)}{P(D=\bullet \mid I)} = \frac{3/5 \cdot 0'25}{9/20} = 1/3$$

TOMA DE DECISIONES

OBJETIVO : Elegir la hipótesis H_i de entre $H_1, H_2, \ldots, H_t$ "a la luz" de los datos D.

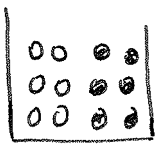
– Máxima probabilidad a posteriori (MAP) + correcta

$$H_i = \max_i P(H_i \mid D) = \max_i \frac{P(D \mid H_i)\, P(H_i)}{P(D)} =$$

$$= \boxed{\max_i P(D \mid H_i)\, P(H_i)}$$

$$\boxed{H_i = \max_i P(D \mid H_i)}$$

<u>Ej 2</u> : (NO CAUSAL) Las probabilidades pueden representar conexiones lógicas, no causales. Consideremos una urna con 6 bolas blancas y 6 bolas negras



· Si en la 1ª extracción se ha eliminado una bola blanca, ¿cuál es la prob. de extraer una bola blanca en la 2ª extracción?

$$P(\text{segunda} = O \mid \text{primera} = O) = 5/11$$

· Por el contrario, supongamos que la 2ª extracción ha sido blanca, ¿cuál es la prob. de que la 1ª fuera blanca también?

APLICANDO BAYES ...

$$P(p = O \mid s = O) = \ldots = \frac{5/11 \cdot 6/12}{P(s=O)} = 5/11$$

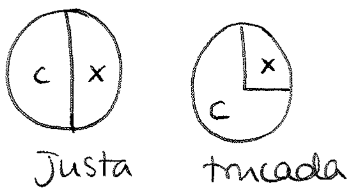
$$P(p = \bullet \mid s = O) = \ldots = \frac{6/11 \cdot 6/12}{P(s=O)} = \boxed{6/11} \text{ Decisión por MAP}$$

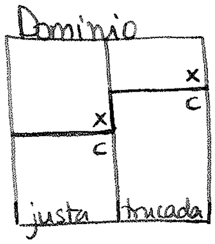
$$P(S=O) = \underbrace{5/11 \cdot 6/12 + 6/11 \cdot 6/12}_{\text{suma de los numeradores}} =$$

<u>Ej 3</u>: Según el hombre del tiempo la prob. de lluvia hoy es del 20%. Estamos en un sótano sin ventanas y no podemos saber qué tiempo hace fuera. Sin embargo, vemos entrar a alguien llevando un paraguas. Sabiendo la prob. de que alguien lleve un paraguas y esté lloviendo es del 70% y solo del 10% en caso contrario. ¿Cuál es la prob. de que esté lloviendo?

$$P(\text{estar lloviendo}) = \frac{0'7 \cdot 0'2}{0'7 \cdot 0'2 + 0'8 \cdot 0'1} = \frac{0'14}{0'22} = 63\%$$

$\boxed{\text{monedas}}$ Saco con monedas justas y trucadas





– Si $D=c$

$$P(H=j \mid D=c) = \frac{P(D=c \mid H=j)\,P(H=j)}{P(D=c)} = \frac{0'5 \cdot 0'5}{P(D=c)} = 0'4$$

$$P(H=t \mid D=c) = \frac{P(D=c \mid H=t)\,P(H=t)}{P(D=c)} = \frac{0'75 \cdot 0'5}{P(D=c)} = 0'6$$

– Si $D=x$

$P(H=j \mid D=x) = 0'67$

$P(H=t \mid D=x) = 0'33$

4 Modelos (o clasificadores)

$D \in \{c, x\} \Longrightarrow H \in \{t, j\}$

modelo h_1	modelo h_2	modelo h_3	modelo h_4
$h_1(c)=t$	$h_2(c)=j$	$h_3(c)=j$	$h_4(c)=t$
$h_1(x)=j$	$h_2(x)=t$	$h_3(x)=j$	$h_4(x)=t$

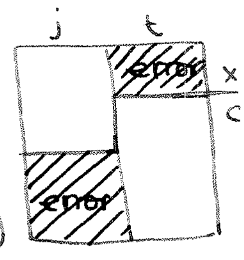
¿Cuál es el mejor modelo? El que tenga menor prob. de equivocarse.

Prob. de error de h_1

$h_1(x)=j$ $\quad P(\text{error}_1) = P(H=j, D=c) +$

$h_1(c)=t$ $\quad P(H=t, D=x) =$

$= P(D=c \mid H=j)\,P(H=j) + P(D=x \mid H=t)\,P(H=t)$

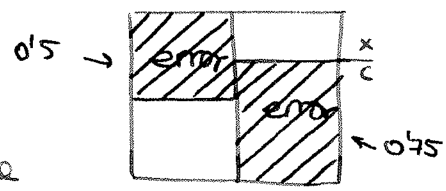


fallo del 37'5%

fallo de h_2

$h_2(c) = j$ $\quad$ $p(\text{error}_2) = 0'5 \cdot 0'5 + 0'75 \cdot 0'5 = 0'625$

$h_2(x) = t$



fallo de h_3 y h_4 = 50%, por lo que el mejor modelo es h_1

REGLA DE BAYES SIMÉTRICA (se coje el modelo que tenga **mayor prob. al aplicar Bayes**)

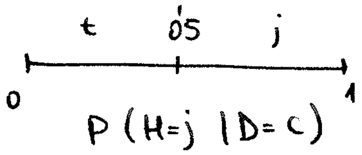
El mejor modelo es el que clasifica cada ejemplo con la clase que tiene la mayor prob. a posteriori.

~~Es~~ decir, si $D=c$ diremos $h(c)=j$ si $P(H=j \mid D=c) > P(H=t \mid D=c)$. Diremos $h(c)=t$ si $P(H=t \mid D=c) > P(H=j \mid D=c)$

$P(H=j \mid D=c) + P(H=t \mid D=c) = 1$ decidimos $h(c)=j$ si

$P(H=j \mid D=c) > 1 - P(H=j \mid D=c)$

$P(H=j \mid D=c) > 1/2$



---

20 Sept

MATRIZ DE CONFUSIÓN

Forma de visualizar el error en tabla por modelos.

- columnas = clase real
- filas = clase predicha por el clasificador.

suma diagonal = aciertos
resto = fallos

M.C. para h_1:

Predicción	REAL: $H=j$ +	REAL: $H=t$ −
$h_1=j$ +	25%	12'5%
$h_1=t$ −	25%	37'5%

positivos verdaderos

falsos positivo

Negativos verdaderos

falsos negativo

REGLA ASIMÉTRICA DE BAYES

- Suponemos un riesgo (corte) R si pasamos t por j
- $\qquad\qquad\qquad\qquad\qquad\quad R'$ $\qquad\qquad\qquad j$ por t

Def: Matriz de corte

	$M=j$	$M=t$
$h=j$	0	R
$h=t$	R'	0

¿Cuál es el corte si $D=c$ y clasificamos como:

$h_*(c)=j \;:\; P(M=t\,|\,D=c)\times R$

$h_*(c)=t \;:\; P(M=j\,|\,D=c)\times R'$

si $D=c$

$$h_*(c)\begin{cases} j & R\times P(M=t\,|\,D=c) < R'\times P(M=j\,|\,D=c) \\ t & \text{en caso contrario}\end{cases}$$

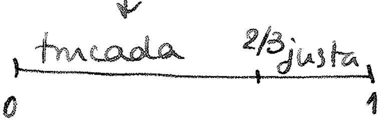
dado que $P(M=t\,|\,D=c)+P(M=j\,|\,D=c)=1$

diremos $h(c)=j$ si $R\times(1-P(M=j\,|\,D=c)) < R\times P(M=j\,|\,D=c)$

$$P(M=j\,|\,D=c) > \frac{R}{R+R'}$$

Ejemplo: $R=2R'$

$$P(M=j\,|\,D=c) > \frac{2R'}{R'+2R'} = \frac{2}{3}$$



Def: Si tenemos dos hipótesis H_1 y H_2 y si al elegir H_1 se corre un riesgo R y si elegimos H_2, se corre un riesgo R', entonces hay que elegir H_1 si $\boxed{P(H_1|D) > \frac{R}{R+R'}}$ y H_2 en el caso contrario

Ejemplo (MONEDAS)

si $R = 3R'$ ¿ h_1, h_2, h_3 ó h_4?

multiplicamos $error_c \cdot$ riesgo de ese error + $error_x \cdot$ su riesgo

$\boxed{\begin{array}{l} h_1(c) = t \\ h_1(x) = j \end{array}}$

$<\text{riesgo } h_1> = R' \cdot P(M=j, D=c) + R\, P(M=t, D=x) =$

$\cancel{= 3R'(0'4 \cdot 0'625) + R'(0'33 \cdot 0'375) = R \cdot 0'625}$

$= R' \cdot P(D=c \mid M=j) P(M=j) + 3R' \cdot P(D=x \mid M=t) P(M=t) =$

$= R' \, 1/4 + 3R' \, 1/8 = 3R' \frac{5}{8} = \underline{\underline{R' \cdot 0'625}}$

* Calcular h_2, h_3 y h_4.

25 Sept '13

$\boxed{\begin{array}{l} h_2(c) = j \\ h_2(x) = t \end{array}}$

$<\text{Riesgo } h_2> = R \cdot p(c,t) + R' \cdot p(x,j) =$

$= R \cdot p(c,t) p(t) + R' \cdot p(x \mid j) P(j) =$

$= 3R' \cdot 0'75 \cdot 0'5 + R' \cdot 0'5 \cdot 0'5 = \underline{\underline{R' \cdot 1'375}}$

$\boxed{\begin{array}{l} h_3(c) = j \\ h_3(x) = j \end{array}}$

$<\text{Riesgo } h_3> = P(h_3=j, t) \cdot R + P(h_3=t, j) R' =$

$= P(c,t) \cdot R + P(x,t) R =$

$= 0'75 \cdot 0'5 \cdot 3R' + 0'25 \cdot 0'5 \, 3R' = \underline{\underline{1'5R'}}$

$\boxed{\begin{array}{l} h_4(c) = t \\ h_4(x) = t \end{array}}$

$<\text{Riesgo } h_4> = P(c,j) R' + P(x,j) R' = \underline{\underline{0'5R'}}$ Hipótesis con la que se corren menos riesgos.

se escoge por la regla asimétrica de Bayes.

¿cumple la regla simétrica de Bayes?

otra manera de hacerlo

x
$$\begin{cases} P(h=t \mid D=x) = 1/3 \quad 1/3 > \frac{R'}{R+R'} = 0'25 \longrightarrow \text{Sí la cumple} \\ P(h=j \mid D=x) = 2/3 > \frac{R}{R+R'} = 0'75 \longrightarrow \text{No lo cumple.} \end{cases}$$

c
$$\begin{cases} P(t \mid c) = 3/5 > \frac{R'}{R'+R} = 0'25 \longrightarrow \text{lo cumple} \\ P(j \mid c) = 2/5 \not> \frac{R}{} = 0'75 \longrightarrow \text{NO lo cumple.} \end{cases}$$

Problema lentillas

$H = n, d, b$ (diagnóstico)

llega paciente con $\boxed{p = (l=m, p=n)}$

$$P(d=n \mid l=m, p=n) = \frac{P(l=m, p=n \mid d=n)\,P(d=n)}{P(p=n, l=m)}$$

Probabilidades a posteriori

$P(d=n) = \frac{15}{24}$; $P(d=d) = 4/24$; $P(d=b) = 5/24$

$P(p=n, l=m \mid d=n) = 1/15$; $P(p=n, l=m \mid d=d) = 3/4$

$P(p=n, l=m \mid d=b) = 2/5$

$$P(d=n \mid p=n, l=m) = \frac{1/15 \cdot 15/24}{P(p=n, l=m)} = 1/6$$

$$P(d=d \mid p=n, l=m) = \frac{3/4 \cdot 4/24}{P(p=n, l=m)} = 1/2 \longleftarrow \text{más probable}$$

$$P(d=b \mid p=n, l=m) = \frac{2/5 \cdot 5/24}{P(p=n, l=m)} = 1/3$$

$\boxed{P(l=h, p=n)}$

	$d=n$	$d=d$	$d=b$
$l=h, p=n$	1/3	1/6	$\boxed{1/2}$
$l=m, p=r$	$\boxed{1}$	0	0
$l=h, p=r$	$\boxed{1}$	0	0

$\square$ Decisión MAP

~~Problema~~

2 Octubre '13

$$P(H_i \mid x_1 \dots x_n) = \frac{P(x_1 \dots x_n \mid H_i)\, P(H_i)}{P(x_1 \dots x_n)} = \frac{P(x_1 \mid H_i) \dots P(x_n \mid H_i)\, P(H_i)}{P(x_1 \dots x_n)}$$

Suposición Naïve Bayes

REGLA DE DECISIÓN NAÏVE BAYES

$$\boxed{\text{CLASE NB} = \underset{H_i}{\arg\max} \prod_j^n \underbrace{P(x_j \mid H_i)}_{\text{Estimado a partir de datos.}} P(H_i)}$$

Ejemplo:

$$P(l=m,\, p=n \mid d=n) \underset{\text{Sup NB}}{=} P(l=m \mid d=n)\, P(p=n \mid d=n) = \frac{7}{15} \times \frac{3}{15} = \frac{7}{75} \neq 1/15$$

$P(l=m, p=n \mid d$

Para calcular la clasificación NB hay que cruzar atributos con clases.

	n	b	d
miopía	7	2	3
hipermetropía	8	3	1

	n	b	d
normal	3	5	4
reducidas	12	0	0

$(l=m,\ p=n)$

$$P(d=n \mid l=m, p=n) = \frac{P(l=m \mid d=n)\, P(p=n \mid d=n)\, P(d=n)}{P(l=m,\ p=n)} = \frac{7/15 \cdot 3/15 \cdot 15/24}{P(\dots)}$$

$$= \frac{7/120}{P(\dots)}$$

$$P(d=b \mid l=m, p=n) = \frac{2/5 \cdot 5/5 \cdot 5/24}{P(\dots)} = \frac{1/12}{P(\dots)}$$

Elegir el mayor numerador

$$P(d=d \mid l=m, p=n) = \frac{3/4 \cdot 4/4 \cdot 4/24}{P(\dots)} = \boxed{\frac{1/8}{P(\dots)}}$$

* El denominador se calcula sumando todos los numeradores.

APROXIMACIÓN	EXACTO
$P(d=u \mid l=m, p=n) = 22\%$	17%
$P(d=b \mid l=m, p=n) = 31\%$	37%
$P(d=d \mid l=m, p=n) = 47\%$ ⇦	50% ⇦

$(l=m, p=r)$

$$P(d=u \mid l=m, p=r) = \frac{7/30}{P(\dots)} = 100\%$$

$$P(d=b \mid l=m, p=r) = \frac{\emptyset}{P(\dots)} = 0\%$$

$$P(d=d \mid l=m, p=r) = \frac{\emptyset}{P(\dots)} = 0\%$$

CORRECCIÓN DE LAPLACE

En el caso de que tengamos valores con ceros en las tablas, se añade 1 a todas las celdas.

	u	b	d
normal	3 +1	5 +1	4 +1
reducida	12 +1	0 +1	0+1

$(l=m, p=r)$ con NB y corrección

Es la prob. original, no de la nueva tabla.

$$P(d=u \mid l=m, p=r) = \frac{7/15 \cdot 13/17 \cdot 15/24}{P(\dots)} = \frac{91/408}{P(\dots)}$$

$$P(d=b \mid l=m, p=r) = \frac{2/5 \cdot 1/4 \cdot 5/24}{P(\dots)} = \frac{1/84}{P(\dots)}$$

$$P(d=d \mid l=m, p=r) = \frac{3/4 \cdot 1/6 \cdot 4/24}{P(\dots)} = \frac{1/48}{P(\dots)}$$

4 oct '13

PROBLEMA FLORES

Naïve Bayes para atributos cuantitativos.

Prob. a posteriori $\quad p(\text{setosa} \mid ap = 1'5) = \emptyset$

~~p(a = 1.5)~~ $\quad p(\text{versicolor} \mid ap = 1'5) = \dfrac{p(ap = 1'5 \mid \text{ver})\, p(\text{vers})}{p(ap = 1'5)}$

* 150 Ejemplos.
  50 por clase

$$= \frac{17/50 \cdot 50/150}{p(ap = 1'5)} = \underline{\underline{0'85}}$$

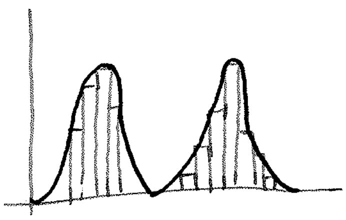
$$p(\text{virgínica} \mid ap = 1'5) = \frac{3/50 \cdot 50/150}{p(ap = 1'5)} = 0'15$$

* Suponemos que la variable aleatoria ap sigue una normal de media μ y varianza σ^2.

Dada una muestra $\{x_1, x_2, \dots, x_n\}$ la verosimilitud viene dada por

$$\boxed{p(x_1, x_2, \dots, x_n \mid \mu \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}$$

Objetivo : hallar la media y la varianza para esta muestra.



* ver que gaussiana se acerca más a los datos.

Simplificando la fórmula aplicando el logaritmo en ambos lados.

$$\ln p(x_1, x_2, \dots, x_n \mid \mu \sigma^2) = \sum_{i=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right)$$

$$'' \quad = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} + \ln e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right) =$$

$$'' \quad = -n \cdot \ln\left(\sqrt{2\pi\sigma^2}\right) + \sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}$$

¿Qué valores de μ y σ^2 maximizan la verosimilitud de la muestra?

$\underline{\underline{\sigma^2}}$

$$\frac{\partial \ln P(x_1, x_2, \dots, x_n \mid \mu \sigma^2)}{\partial \sigma^2} = n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{2\pi}{2}(2\pi\sigma^2)^{-1/2} +$$

$$+ \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

para encontrar el máximo.

$$\frac{n \cdot 2\pi}{4\pi\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$n + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

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

$$\boxed{\mu = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

← se demuestra derivando la ecuación principal respecto a μ ~~y~~ e igualando a cero para maximizar.

<u>ap</u>

	setosa	versicolor	virgínica
μ	0'246	1'326	2'026
σ^2	0'0111	0'0391	0'0754

ap = 1'5

$$p(set \mid ap = 1'5) = \frac{P(ap = 1'5 \mid set) \, p(set)}{p(ap = 1'5)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2_{set}}} e^{-\frac{(1'5 - \mu_{set})^2}{2\sigma^2_{set}}} \cdot P(set)}{P(ap = 1'5)}$$

9 octubre 2013

Ejemplo 2.

$ls = 6'3$ $\quad$ $p(set \mid \dots) = \emptyset$

$ap = 1'6$ $\quad$ $p(\dots \mid ver)\, p(ver) = 0'124$

$p(\dots \mid vir)\, p(vir) = 0'07$

* MODELOS GRÁFICOS (Redes de Bayes)

Necesitamos la regla de la cadena: Aplicación de la regla del producto múltiples veces.

Considerando las variables aleatorias d, e, l, p

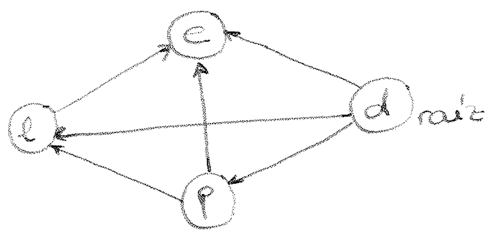
$$p(e, l, p, d) = p(e \mid l, p, d) \cdot p(l, p, d) =$$

$$= p(e \mid l, p, d) \cdot p(l \mid p, d) \cdot p(p, d)$$

$$= p(e \mid l, p, d)\, p(l \mid p, d) \cdot p(p \mid d) \cdot p(d)$$

Representación gráfica

(mediante grafo acíclico dirigido)



GRAFO GENERAL:

$$P(x_1 \dots x_n) = \prod_1^n p(x_i \mid pa_i)$$

$pa_i \equiv$ conjunto de los padres de la variable x_i.

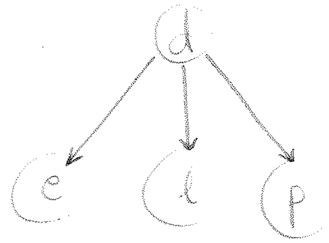
Si x_i no tiene padres $p(x_i \mid pa_i) = p(x_i)$

NOTA: $$p(d \mid l, e, p) = \frac{p(l, e, p \mid d)\, p(d)}{p(l, e, p)} = \frac{p(l, e, p, d)}{p(l, e, d)}$$

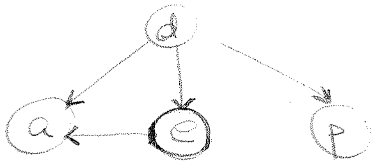
NB ⇒ $$p(l, e, p, d) = p(e \mid l, p, d)\, p(l \mid p, d)\, p(p \mid d)\, p(d) =$$

$$= p(e \mid d)\, p(l \mid d)\, p(p \mid d)\, p(d)$$

Por Naive Bayes los nodos solo dependen de la clase.



- Ejemplo de las lentillas



calcular: $p(d \mid a,e,p) =$

$$= \frac{P(a,e,p,d)}{p(a,e,p)} =$$

'a' depende de 'e' y 'd'

$$= \frac{P(a \mid e,d)\, p(e \mid d)\, p(p \mid d)\, p(d)}{p(a,e,p)}$$

'd' no depende de nada.

Ahora podemos asignar valores a las variables

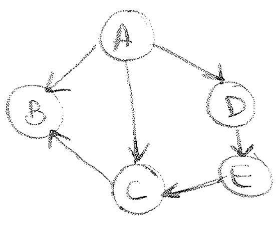
$D \{ e: \text{joven},\ p = \text{normal},\ a = \text{sí} \}$ // $H = d \in \{n, d, b\}$

$d = NO \longrightarrow P(d=n \mid a=sí, e=j, p=n) = \dfrac{\frac{2}{4} \cdot \frac{4}{15} \cdot \frac{3}{15} \cdot \frac{15}{24}}{p(\dots)} = \dfrac{1/60}{p(\dots)}$

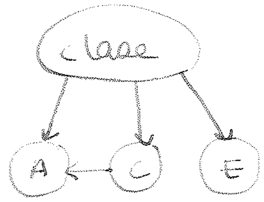
$d = \text{duras} \longrightarrow P(d=d \mid e=j, p=n, a=sí) = \dfrac{2/2 \cdot 2/4 \cdot 4/4 \cdot 4/24}{p(\dots)} = \dfrac{1/12}{P(\dots)}$

Ejemplo: Calcular la verosimilitud de B, C, D y E dado A

$p(b \mid a,c) \cdot p(c \mid a,e) \cdot P(d \mid a) \cdot p(e \mid d)$



Ejemplo :



Tenemos $A=1$ $C=0$ $E=0$

¿ Clase más probable?

A	B	C	D	E	clase
1	0	1	0	0	A
1	1	0	0	1	A
0	1	0	1	0	S
0	1	1	1	1	S
0	0	1	0	1	A
1	0	1	0	0	A
1	1	1	1	0	A
0	0	0	1	1	S
1	0	0	0	1	A
1	1	0	1	0	S

$p(a,c,e,clase) =$

$$= \frac{p(a|c,cl)\,p(c|cl)\,p(e|cl)\,p(cl)}{P(\ldots)}$$

Solución: 67% clase A

33% clase S

11 octubre 2013

VECINOS PRÓXIMOS (K-NN)

Hasta ahora la prob de la verosimilitud

. Discretos : frecuencias

$$P(x=v|h_1) = \frac{\#\text{ casos con } x=v \text{ y } h_1}{\#\text{ casos de } h_1}$$

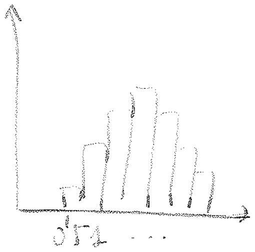
. Cuantitativos :

a) Discretizar ; generar histogramas y usar frecuencias. Estimación no paramétrica

b) Estimar la densidad de prob. (con gaussiana) Estimación paramétrica.

~~1. Por histogramas~~

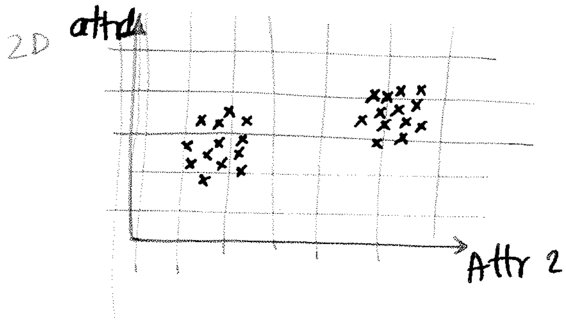
función prob. de densidad = pdf (En nuestro caso pdf = gaussiana param μ y σ^2

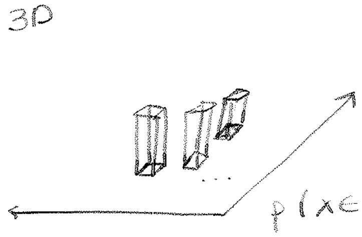


$$p(1'4 < x < 1'5) = \int_{1'4}^{1'5} pdf(x)\, dx$$

* No siempre se puede aplicar una gaussiana.

Vamos a generalizar la estimación por histogramas a D-dimensiones.





$$p(x \in Región) = \int_{Región} pdf(\bar{x})\, d\bar{x}$$

Estimación de:

$$p(\bar{x} \in Región) = \int_{Región} pdf(\bar{x})\, d\bar{x}$$

1: Como histogramas con frecuencias.
Sean K_R el número de puntos dentro de la región R y N el número total de puntos.

$$P(x \in R) \simeq K_R / N$$

N → Debe de ser grande para que la aproximación sea válida y el espacio tiene que estar bien cubierto de puntos.

2: Si ponemos R pequeña, podemos suponer que $pdf(\bar{x})$ es cte. en dicha región y

$$p(\bar{x} \in R) = \underbrace{pdf(\bar{x}_R)}_{altura} \cdot \underbrace{\int_R d\bar{x}}_{\text{Base o volumen. Depende de la dimensión}} = pdf(\bar{x}_R) \cdot \underbrace{V_R}_{volumen}$$

* Igualando 1 y 2:

$$p(\bar{x} \in R) = \frac{k_R}{N} = pdf(\bar{x}_R)V_R$$

$$\boxed{pdf(\bar{x}_R) = \frac{k_R}{NV_R}}$$

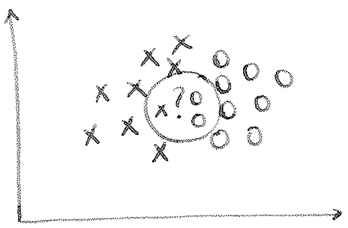
* con N suficientemente grande y V_R suficientemente pequeño

– FORMAS DE USAR LA FÓRMULA

- Fijando k_R (número de puntos en R) y obteniendo el V_R necesario para que haya k_R puntos
  – VECINOS PRÓXIMOS (K-NN) –

- Fijando V_R y calculando k.
  – MÉTODOS DE NÚCLEOS (kernel) –

Ejemplo: Fijamos k_R



x : clase 1
o : clase 2

$k_R = 4$
$N = 15$; $N_1 = 8$; $N_2 = 7$
? : pto. de consulta.

16 octubre 2013

$$p(\bar{x}_R) = \frac{k_R}{NV_R}$$

→ Estimación de densidad de probs.

- verosimilitud

$$p(\bar{x} \mid C_c) = \frac{k_c}{N_c V_R}$$

k_c es el número de datos de la clase C_c que hay en V_R.

- Evidencia

$$p(\bar{x}) = \frac{k}{NV_R}$$

- Prob. a priori

$$P(C_c) = \frac{N_c}{N}$$

Si aplicamos Bayes a $P(C_c|\bar{x})$ obtenemos la prob. a posteriori.

$$P(c_c|\bar{x}) = \frac{P(\bar{x}|c_c)\,P(c_c)}{P(\bar{x})} = \frac{\frac{\cancel{K}_c}{\cancel{N_c V_R}} \; \frac{N_c}{\cancel{N}}}{\cancel{K}/N V_R} = \frac{k_c}{k}$$

$$\boxed{P(c_c|\bar{x}) = \frac{k_c}{k}}$$

... aplicando la fórmula al ejemplo...

$P(o|?) = 3/4 \longrightarrow$ con $k = 4$

$P(x|?) = 1/4$

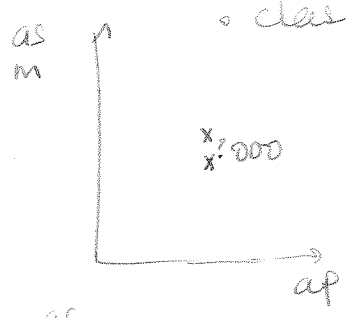
Regla de clasificación k-NN

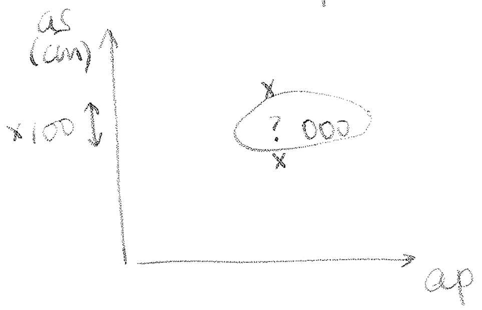
<u>Def</u>: La clase más común dentro de los k vecinos más próximos al punto a clasificar

ALGORITMO k-NN

- Entrenamiento
  → No hay, se guardan todos los ejemplos.

- clasificación $(\bar{x})$
  → calcular las distancias al resto de ejemplos con métricas.
  → seleccionar los k más próximos
  → Devolver el más común.



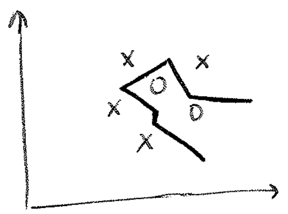


Solución habitual a la métrica: Distancia euclídea y normalizar cada atr para que tenga media 0 y varianza 1.

Selección del k óptimo.

- calculamos la matriz de distancias entre todos los puntos de entrenamiento
- para cada pto. clasificar probando todos los k y comparar con la clase real.
- Resolver k con menor error.

Frontera de clasificación.



$k = 1$

Polígonos de Thiessen o diagrama de voronoi.

* Otros problemas de k-NN

- existencia de attr. irrelevantes.
- coste
  - En memoria
  - En tiempo de clasificación

* Selección de prototipos: Buscar un subconjunto S del original T.

ALGORITMO: CNN (condensed NN)

- Inicializamos S con un ejemplo por clase al azar.
- do (para cada ejemplo x en T
  - si la clase asignada a $\bar{x}$ por S es incorrecta
    $S = S \cup \bar{x}$

) while (!cambie S)

Para attr. discretos: la distancia es 1 a los que sean distintos

$$dist(\bar{x}_1, \bar{x}_2) = \sum_{1}^{n\ attr} I(\bar{x}_1[i] \neq \bar{x}_2[i])$$

$I(true) = 1$

$I(false) = \emptyset$

18 octubre '13

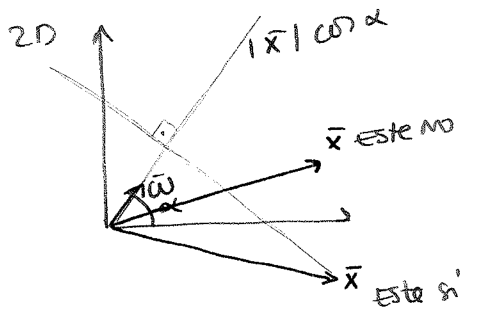
Clasificadores lineales

Clasificadores cuya frontera de decisión es un plano.

Sea h un clasificador lineal:

$$h(\bar{x}) = \begin{cases} C_1 & \text{si } \bar{\omega}^T \cdot \bar{x} + \omega_0 \geq \emptyset \\ C_2 & \text{si } \bar{\omega}^T \cdot \bar{x} + \omega_0 < \emptyset \end{cases}$$

ω_0 es una cte. $\bar{\omega}$ = vector de D-dimensiones



La frontera son los ptos de x que cumplen:

$$\bar{\omega}^T \cdot \bar{x} + \omega_0 = \emptyset$$

$$|\bar{\omega}||\bar{x}| \cos\alpha + \omega_0 = \emptyset$$

$$|\bar{x}| \cdot \cos\alpha = \frac{-\omega_0}{|\bar{\omega}|}$$

23 octubre '13

Inciso

FUNCIÓN SIGMOIDAL

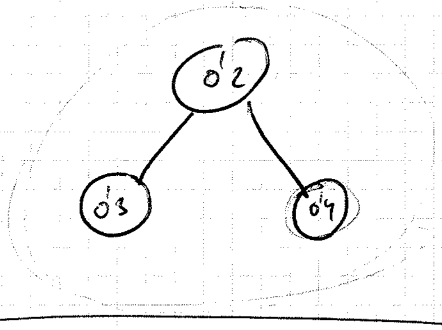
Considerando un problema de 2 clases C_1 y C_2 la prob. a posteriori es:

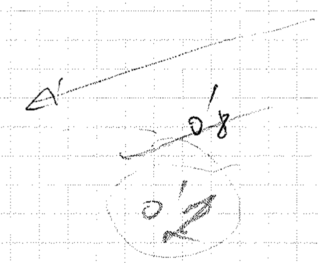
$$P(C_1|\bar{x}) = \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x})} = \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_1)P(C_1) + P(\bar{x}|C_2)P(C_2)}$$

Alumno........ Curso........ Grupo........
Asignatura........
día........ mes........ de........
EXAMEN DE EVALUACIÓN

$$X_i = \frac{X_i - \mu}{\sigma}$$

K=3





Selección del K óptimo mediante validación en train:

K_optimo()

validación leave one out

Se calcula distancia entre todos los pares de puntos
Para cada ejemplo en train
Se ordenan el resto de ejemplos por distancias
Se hacen tantas clasificaciones como valores de K se están probando
Se calcula el error para cada K y devolvemos el K con menor error.

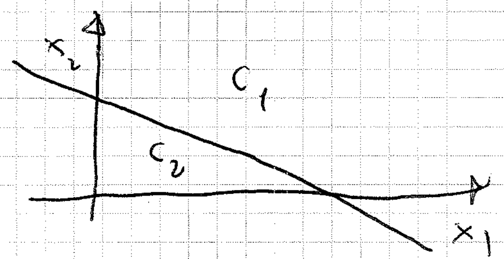
Más inconvenientes

- Existencia de atributos irrelevantes → produce distancias que no son relevantes
- Coste → para cada ejemplo a clasificar hay que calcular tantas distancias como ejemplos en train.

clasificador lineal

vector perpendicular a la recta

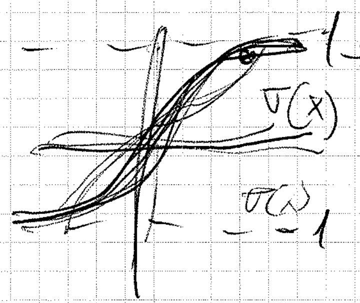
$$h(\bar{x}) = \begin{cases} C_1 & \text{si } \bar{\omega}^T \bar{x} + \omega_0 > 0 \\ C_2 & \text{si } \bar{\omega}^T \cdot \bar{x} + \omega_0 < 0 \end{cases}$$



frontera $\Rightarrow$ $\bar{\omega}^T \cdot \bar{x} + \omega_0 = 0$

$$|\bar{\omega}^T| \, |\bar{x}| \cos\alpha + \omega_0 = 0$$

$$|\bar{x}| \cos\alpha = \frac{-\omega_0}{|\bar{\omega}|}$$

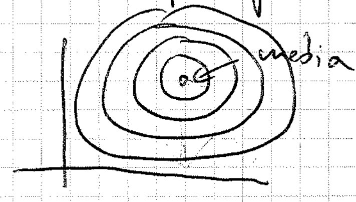


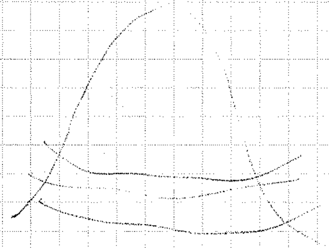
Vamos a ver cómo surgen de forma "natural" haciendo suposiciones sencillas sobre los datos

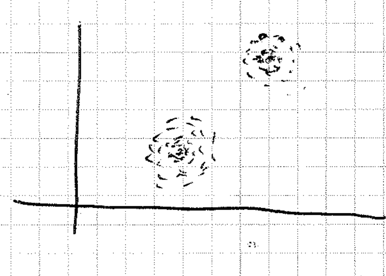
Matriz de covarianzas $\Rightarrow$ identidad

$\hookrightarrow$ los puntos están desparramados por igual

$\hookrightarrow$





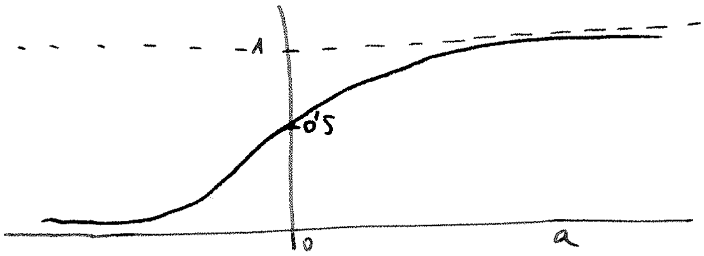


...dividimos el denominador entre $P(\bar{x}|C_1)P(C_1)$

$$P(C_1|\bar{x}) = \frac{1}{1+\frac{P(\bar{x}|C_2)P(C_2)}{P(\bar{x}|C_1)P(C_1)}} = \frac{1}{1+e^{-a}} = \sigma(a)$$

con $a = \ln \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_2)P(C_2)}$

Función sigmoidal o de activación.



Decisión:

$$\begin{cases} \text{si } P(C_1|x) > 0'5 \longrightarrow C_1 \\ \text{si no} \longrightarrow C_2 \end{cases}$$

$$\begin{cases} \text{si } a > \emptyset \longrightarrow C_1 \\ \text{si no} \longrightarrow C_2 \end{cases}$$

Suponemos que la verosimilitud sigue una distribución normal $N(\bar{\mu}_k, \Sigma_k)$ dada la clase k.

$$P(\bar{x}|C_k) = N(\bar{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu}_k)^T \cdot \Sigma_k^{-1}(\bar{x}-\bar{\mu}_k)\right\}$$

vector de medias ($\bar{\mu}_k$) — matriz de covarianzas (Σ_k) — dimensiones (D)

Caso 1: Suponemos dos clases y $\Sigma_k = \mathbb{I}$ Identidad.

$$a = \ln \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_2)P(C_2)} = \ln\left[\exp\left\{\frac{(\bar{x}-\bar{\mu}_1)^2}{2} - \frac{(\bar{x}-\bar{\mu}_2)^2}{2}\right\}\frac{P(C_1)}{P(C_2)}\right] =$$

$$= \frac{(\bar{x}-\bar{\mu}_1)^2}{2} - \frac{(\bar{x}-\bar{\mu}_2)^2}{2} + \ln\frac{P(C_1)}{P(C_2)} > \emptyset \Rightarrow C_1$$

<u>Caso 1.1.</u> : Si las prob. a priori son iguales $P(C_1) = P(C_2)$

$$\longrightarrow \quad \frac{(\bar{x} - \bar{\mu}_2)^2}{2} > \frac{(\bar{x} - \bar{\mu}_1)^2}{2}$$

· si · $(\bar{x} - \mu_2{}^2) > (\bar{x} - \bar{\mu}_1)^2 \longrightarrow C_1$

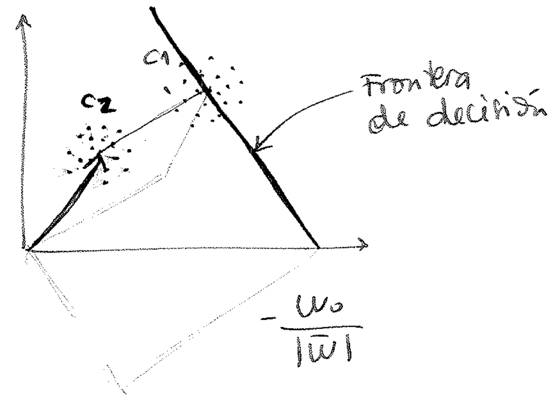
¿Qué frontera tenemos?

$P(C_1|\bar{x}) = P(C_2|\bar{x}) = 0'5$

$a = \emptyset$

$$\frac{(\bar{x} - \bar{\mu}_2)^2}{2} - \frac{(\bar{x} - \bar{\mu}_1)^2}{2} - \frac{\bar{\mu}_1^2 + \bar{x}^2 - 2\bar{x}\mu_1}{2} + \ln \frac{P(C_1)}{P(C_2)} = \emptyset$$

$$(\bar{\mu}_1 - \mu_2)^T \bar{x} + \frac{1}{2}\left(\bar{\mu}_2^2 - \bar{\mu}_1^2\right) + \ln \frac{P(C_1)}{P(C_2)} = \emptyset$$



$\bar{w}^T \cdot \bar{x} + w_0 = \emptyset$

$\bar{w} = \bar{\mu}_1 - \bar{\mu}_2$

$w_0 = \frac{1}{2}\left(\bar{\mu}_2^2 - \bar{\mu}_1^2\right) + \ln \frac{P(C_1)}{P(C_2)}$

¿Qué pasa si $P(C_1) > P(C_2)$? ¿Hacia dónde se ~~mueve~~ mueve el hiperplano?

<u>Caso 2</u>: Suposiciones
- 2 clases
- $\Sigma = \Sigma_1 = \Sigma_2$

$$a = \ln \frac{P(\bar{x}|C_1)\,P(C_1)}{P(\bar{x}|C_2)\,P(C_2)} = \ln\left[\exp\left\{\frac{-1}{2}(\bar{x} - \bar{\mu}_1)\Sigma^{-1}(\bar{x} - \bar{\mu}_1) + \frac{1}{2}(\bar{x} - \bar{\mu}_2)\Sigma^{-1}(\bar{x} - \bar{\mu}_2)\right\} \cdot \frac{P(C_1)}{P(C_2)}\right] =$$

$$= -\frac{1}{2}(\bar{x}-\bar{\mu}_1)\Sigma^{-1}(\bar{x}-\mu_1) + \frac{1}{2}(\bar{x}-\mu_2)\Sigma^{-1}(\bar{x}-\mu_2) + \ln\frac{P(C_1)}{P(C_2)} =$$

$$= -\frac{1}{2}(\bar{x}\Sigma^{-1}\bar{x} - \mu_1\Sigma^{-1}\bar{x} - \bar{x}\Sigma^{-1}\bar{\mu}_1 + \bar{\mu}_1\Sigma^{-1}\bar{\mu}_1) +$$

$$+ \frac{1}{2}(\bar{x}\Sigma^{-1}\bar{x} - \bar{\mu}_2\Sigma^{-1}\bar{x} - \bar{x}\Sigma^{-1}\bar{\mu}_2 + \bar{\mu}_2\Sigma^{-1}\bar{\mu}_2) + \ln\frac{P(C_1)}{P(C_2)} =$$

$$= \frac{1}{2}\left(\left((\bar{\mu}_1-\bar{\mu}_2)\Sigma^{-1} + \left(\Sigma^{-1}(\bar{\mu}_1-\bar{\mu}_2)\right)^T\right)\bar{x} + \frac{1}{2}\left(\bar{\mu}_1\Sigma^{-1}\bar{\mu}_1 + \bar{\mu}_2\Sigma^{-1}\bar{\mu}_2\right) +\right.$$

$$+ \ln\frac{P(C_1)}{P(C_2)}$$

$$\longrightarrow \bar{\omega} = \Sigma^{-1}(\bar{\mu}_1 - \bar{\mu}_2) \; ; \; \omega_0 = -\frac{1}{2}\bar{\mu}_1\Sigma^{-1}\bar{\mu}_1 + \frac{1}{2}\bar{\mu}_2\Sigma^{-1}\bar{\mu}_2 + \ln\frac{P(C_1)}{P(C_2)}$$

Que las matrices de covarianza sean iguales, quiere decir que las nubes de puntos van a tener la misma forma

$$P(C_1|\bar{x}) = \sigma(a) = \sigma(\bar{\omega}^T\bar{x} + \omega_0)$$

$$\ln\frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_2)P(C_2)} = \bar{\omega}^T\bar{x} + \omega_0 \longrightarrow$$ bajo ciertas suposiciones ~~es~~ una recta.

$$[\text{si } P(C_1|\bar{x}) > 0'5] \equiv [\text{si } a > \emptyset] \equiv [\text{si } \bar{\omega}^T\bar{x} + \omega_0 > \emptyset]$$

25 Octubre '13

REGRESIÓN LOGÍSTICA (¡No es regresión! ¡es clasificación!)

Hemos visto que bajo ciertas condiciones $\begin{cases} P(\bar{x}|C_1) = N(\bar{\mu}_1, \Sigma) \\ P(\bar{x}|C_2) = N(\bar{\mu}_2, \Sigma) \end{cases}$

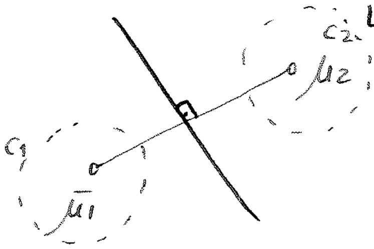
entonces la prob. a posteriori de C_1 es $P(C_1|\bar{x}) = \sigma(\bar{\omega}^T\bar{x} + \omega_0)$

con $\bar{\omega}^T = \Sigma^{-1}(\bar{\mu}_1 - \bar{\mu}_2)$

$$\omega_0 = -\frac{1}{2}\bar{\mu}_1^T\Sigma^{-1}\bar{\mu}_1 + \frac{1}{2}\bar{\mu}_2^T\Sigma^{-1}\bar{\mu}_2 + \ln\frac{P(C_1)}{P(C_2)}$$

Inciso

$\Sigma_1 = \Sigma_2 = \mathbb{I}$



Nube de ptos. con mayor prob. de puntos en el centro.

OP1: Estimar $\bar{\mu}_1$, $\bar{\mu}_2$ y Σ por MV para D-dimensiones

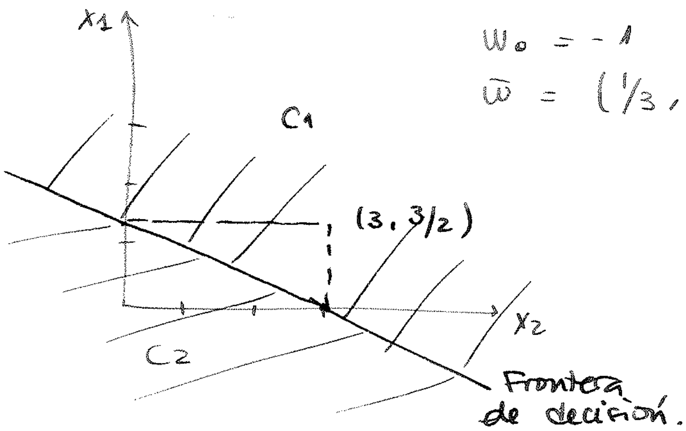
Nº de parámetros: $\underset{\bar{\mu}_1}{D} + \underset{\bar{\mu}_2}{D} + \underset{\Sigma}{\frac{D(D+1)}{2}} + 2$

OP2: Estimar $\bar{w}$ y w_0 por MV $\longrightarrow$

$\underset{D}{\bar{w}} \quad \underset{1}{w_0}$ — $D + 1$ parámetros

Simplificando notación:
$\bar{w} \to (w_0, w_1, w_2, \dots, w_D)$
y los ejemplos
$\bar{x} \to (1, x_1, x_2, \dots, x_D)$
$\Rightarrow P(C_1|\bar{x}) = \sigma(\bar{w} + \bar{x})$

$\longrightarrow$ Ejemplo:



$w_0 = -1$
$\bar{w} = (1/3, 2/3)$

$\Big\} \; -1 + \frac{1}{3}x_1 + \frac{2}{3}x_2 = \emptyset$

$x_2 = \emptyset \; ; \; x_1 = 3$

$x_1 = \emptyset \; ; \; x_2 = 3/2$

¿Prob. de que $(3, 3/2)$ sea C_1?

$$P(C_1|\bar{x}) = \sigma(\underbrace{\bar{w}^T \bar{x} + w_0}_{a}) = \sigma(1) = \frac{1}{1+e^{-1}} = \underline{\underline{0'73}}$$

$$a = \frac{1}{3} \cdot 3 + \frac{2}{3} \cdot \frac{3}{2} - 1 = 1$$

$$x = (0,0) \longrightarrow P(C_1|\bar{x}) = \sigma(-1) = \frac{1}{1+e^{1}} = 0'27.$$

Tenemos un conjunto de entrenamiento

$$D = \{ \{\bar{x}_i, t_i\}, \ i = 1, \dots, N \} \quad \text{donde } t_1 \begin{cases} 1 & \text{si } \bar{x}_i \in C_1 \\ 0 & \text{si } \bar{x}_i \in C_2 \end{cases}$$

En reg. logística...

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad \begin{aligned} H &= \bar{\omega} \\ D &= t_1, \dots, t_N \\ I &= \bar{x}_1, \dots, \bar{x}_N \end{aligned}$$

y con verosimilitud

$$P(t_1, \dots, t_N | \omega, \bar{x}_1 \dots \bar{x}_N) = \prod_{i=1}^{N} P(t_1 | \bar{w}_0, \bar{x}_i) =$$

$$= \prod_{i=1}^{N} \sigma^{t_i} (1 - \sigma_i)^{1-t_i}$$ aplicamos ln para simplificar

$$\ln P(t_1, \dots, t_N | \omega_1, \bar{x}_1 \dots \bar{x}_N) = -\sum_{i=1}^{N} t_i \ln \sigma_i -$$

$$- \sum_{i=1}^{N} (1 - t_i) \ln (1 - \sigma_i)$$

Optimizamos...

30 oct '13

$$\frac{\partial}{\partial \bar{\omega}} \left(- \ln p(t_1 \dots t_N | \bar{\omega} x_i \dots x_N) \right) =$$

$$\frac{\partial E}{\partial \bar{\omega}} = 0; \quad \frac{\partial E}{\partial \bar{\omega}} = \left(- \sum_{i=1}^{N} t_i \frac{1}{\sigma_i} + \sum_{i=1}^{N} (1 + t_i) \frac{1}{1 - \sigma_i} \right) \frac{\partial \sigma_i}{\partial \bar{\omega}} =$$

$$= \left(- \sum_{i=1}^{N} t_i (1 - \sigma_i) + \sum_{i=1}^{N} (1 - t_i) \sigma_i \right) \bar{x}_i = \sum_{i=1}^{N} (- t_i + t_i \sigma_i + \sigma_i - t_i \sigma_i) \bar{x}_i =$$

$$= \sum_{i=1}^{N} (\sigma_i - t_i) \bar{x}_i = \sum_{i=1}^{N} \left(\sigma(\bar{\omega}^T \cdot \bar{x}_i) - t_i \right) \bar{x}_i = 0$$

$$\frac{\partial \sigma_i}{\partial \bar{\omega}} = \frac{\partial}{\partial \omega} \frac{1}{1 + e^{-\bar{\omega}^T \bar{x}_i}} = \frac{+ e^{-\bar{\omega}^T \bar{x}_i}}{(1 + e^{-\bar{\omega}^T \bar{x}_i})^2} = \frac{1}{1 + e^{-\bar{\omega}^T \bar{x}_i}} \cdot \frac{-e^{-\bar{\omega}^T \bar{x}_i}}{1 + e^{-\bar{\omega}^T \bar{x}_i}}$$

$$= \sigma_i (1 - \sigma_i) \bar{x}_i$$

— Para un ejemplo:

$$\frac{\partial E_i}{\partial \bar{w}} = \left(\sigma(\bar{w}^T \bar{x}_i) - t_i\right)\bar{x}_i$$

Para aumentar la verosimilitud del ejemplo x_i hay que mover la recta $\bar{w}$ en sentido opuesto al gradiente $(\sigma_i - t_i)\bar{x}_i$ y proporcional a una cte. de aprendizaje η.

$$\bar{w}_{t+1} \longleftarrow \bar{w}_t - \eta\,(\sigma_i - t_i)\,\bar{x}_i$$ io.

Algoritmo iterativo

cte. entrenamiento (2)

≥ 30

reg_logistica (η, nepocas, datos_train)

- generamos un $\bar{w}$ aleatorio con valores en $[-0'5, 0'5]$ (random) ↳ ArrayList

(los dimensiones se determinan según el nº de atributos + 1)

- para $i_e = 1 \longrightarrow$ nepocas
  - para $i = 1 \longrightarrow N$ // nº de ejemplos → n datos train.
    - calculamos $\sigma(\bar{w}^T \bar{x}_i)$ → sigmoidal.
    - $\bar{w} = \bar{w} - \eta\,(\sigma_i - t_i)\,\bar{x}_i$ clase
      ↳ sigmoidal

return $\bar{w}$.
↳ vector que define el hiperplano.

Ejemplo:

x_0	x_1	x_2	t
1	0	0	0
1	1	0	0
1	0	1	0
1	1	1	1

$\eta = 1$ Generamos un vector para el hiperplano

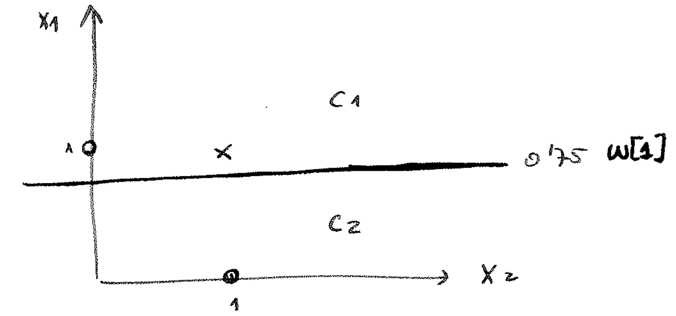
$$\bar{w} = \{\underset{w_0}{-0'3},\ \underset{w_1}{0},\ \underset{w_2}{0'4}\}$$

$$w_0 + x_1 w_1 + x_2 w_2 = 0$$

$$0'3 + 0'4\,x_2 = 0$$

$$\underline{x_2 = 0'75}$$

Para saber a qué clase pertenece calculamos la sigmoidal.



$ie = 1$

$i = 1$

$\sigma(\bar{w}^T x_1) = \sigma(-0'3) =$

$= \dfrac{1}{1+e^{0'3}} = 0'43$

$\bar{w}^T \bar{x}_i = (-0'3, 0, 0'4)\begin{pmatrix}1\\0\\0\end{pmatrix} = -0'3$

$t = 0 \longrightarrow c_2$

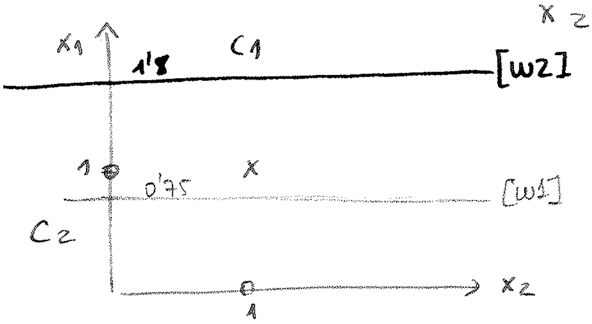
$t = 1 \longrightarrow c_1$

Ahora hay que actualizar el vector $\bar{w}$: $\bar{w} = \begin{pmatrix}-0'3\\0\\0'4\end{pmatrix} - 1 \cdot (0'43 - 0)\begin{pmatrix}1\\0\\0\end{pmatrix}$

$= \begin{pmatrix}-0'73\\0\\0'4\end{pmatrix}$ Significa que la nueva recta frontera o el nuevo vector es igual a:

$-0'73 + x_2 \cdot 0'4 = \emptyset$

$x_2 = \dfrac{-0'73}{0'4} = 1'8$



Iteramos...

$ie = 1$

$i = 2$

$\sigma(\bar{w}^T \bar{x_2}) = \sigma(-0'73) = \dfrac{1}{1+e^{0'73}} = 0'335$

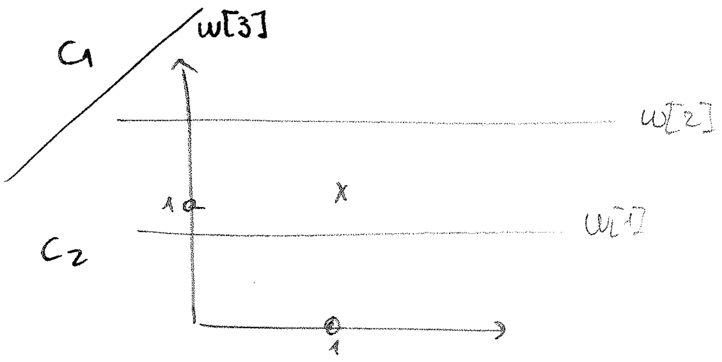
$\bar{w}^T . \bar{x}_2 = (-0'73\ , 0\ , 0'4)\begin{pmatrix}1\\1\\0\end{pmatrix} = \underline{\underline{-0'73}}$

Redefinimos el vector $\bar{w}$

$\bar{w} = \begin{pmatrix}-0'73\\0\\0'4\end{pmatrix} - 1 \cdot (0'335 - 0)\begin{pmatrix}1\\1\\0\end{pmatrix} = \begin{pmatrix}-0'73\\0\\0'4\end{pmatrix} - \begin{pmatrix}0'335\\0'335\\0\end{pmatrix} = \begin{pmatrix}-1'0\\-0'33\\0'4\end{pmatrix}$

$-1'05 - 0'335 x_1 + 0'4 x_2 = \emptyset$

$x_1 = \emptyset\ ;\ x_2 = \dfrac{1'05}{0'4} = 2'6 \ //\ x_2 = \emptyset\ ;\ x_1 = \dfrac{-1'05}{0'33} = \underline{\underline{-3}}$



Iteramos

$ie = 1$

$i = 3$

$$\sigma(\bar{w}^T \cdot \bar{x}_3) = \sigma(-0'05) = 0'34 \longrightarrow \bar{w} = \begin{pmatrix} -1'39 \\ -0'33 \\ 0'06 \end{pmatrix} \longrightarrow$$

$$\longrightarrow -1'39 - 0'33\, x_1 + 0'06\, x_2 = \emptyset$$

$x_1 = \emptyset \longrightarrow x_2 =$

$x_2 = \emptyset \longrightarrow x_1 = \frac{-1'39}{0'33}$

Iteramos

$ie = 1$

$i = 4$

$$\sigma(\bar{w}^T \bar{x}_4) = \sigma(-1'66) = 0'16$$

$$\bar{w}^T \bar{x}_4 = (-1'39, -0'33, 0'66) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = -1'66$$

$$\bar{w} = \begin{pmatrix} -1'34 \\ -0'33 \\ 0'06 \end{pmatrix} - 1\,(0'16 - 1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0'55 \\ 0'5 \\ 0'9 \end{pmatrix}$$

$$-0'55 + 0'5\, x_1 + 0'44\, x_2 = \emptyset$$

$x_1 = \emptyset \longrightarrow x_2 = \frac{0'55}{0'9}$; $x_2 = \emptyset \longrightarrow x_1 \simeq 1'0$

tras 4 épocas : $w = (-2'07 \quad 1'557 \quad 1'944)$

" 12 " : $w = (-3'68 \quad 2'82 \quad 2'82)$

" 100 " : $w = (-8'94 \quad 5'88 \quad 5'94)$

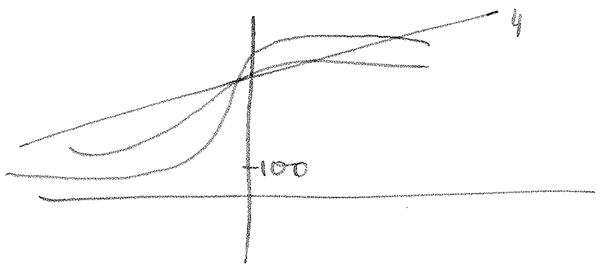
$$P(C_1 | \bar{x}) = \sigma(\bar{w}^T \bar{x})$$

x_0	x_1	x_2	t	4	12	100
1	0	0	0	0'4	0'025	0'00013
1	1	0	0	0'37	0'23	0'045
1	0	1	0	0'47	0'30	0'047
1	1	1	1	0'81	0'84	0'95

$$\sigma(\bar{w}^T \bar{x}_1) = \sigma(-2'07) = 0'11$$

$$\bar{w}^T \bar{x} = \begin{pmatrix} -2'07 & 1'557 & 1'944 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = -2'07$$

¡Sobreaprendizaje!



Reg. logística por MAP

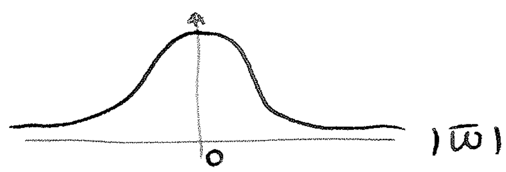
Queremos obtener el máximo de:

$$P(\bar{w} | t_1 \ldots t_N, \bar{x}_1 \ldots \bar{x}_N) \propto \underbrace{\prod_{i=1}^{N} \sigma_i^{t_i} (1-\sigma_i)^{1-t_i}}_{P(t_1 \ldots t_N | \bar{w}, \bar{x}_1 \ldots \bar{x}_N)} \cdot P(\bar{w} | \bar{x}_1 \ldots$$

Proporcional a

Vamos a poner una gaussiana al mod. del vector

$$P(\bar{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\bar{w}^2}{2\sigma^2}}$$



Aplicamos ln

$$E = \ln P(\bar{w} | t_1 \ldots t_N, \bar{x}_1 \ldots \bar{x}_N, \bar{w}) \propto$$

$$\propto -\sum_{1}^{N} t_i \ln \sigma_i - \sum_{1}^{N} (1-t_i) \ln (1-\sigma_i) + \frac{\bar{w}^2}{2\sigma^2}$$

varianza del módulo de $\bar{w}$.

$$\sigma \neq \sigma_i$$

~~Definimos prob. a priori para $\bar{w}$~~

$$\cancel{p(\bar{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|\bar{w}|^2}{2\sigma^2}}}$$

~~La prob. posterior~~

$$\cancel{p(\bar{w} | t_1 \dots t_N \; \bar{x}_1 \dots \bar{x}_N) \propto}$$

6 Noviembre

$$E = \sum_{i=1}^{N} E_i = -\sum_{i=1}^{N} t_i \ln \sigma_i - \sum_{i=1}^{N} (1-t_i) \ln (1-\sigma_i) + \sum_{i=1}^{N} \frac{w^2}{2\sigma^2 N}$$

derivando...

$$\frac{\partial E_i}{\partial \bar{w}} = (\sigma_i - t_i)\bar{x}_i + \frac{\bar{w}}{N\sigma^2} \rightarrow \text{Regla de actualización}$$

$$\bar{w} \leftarrow \bar{w} - \eta \left((\sigma_i - t_i)\bar{x}_i - \frac{\bar{w}}{N\sigma^2} \right)$$

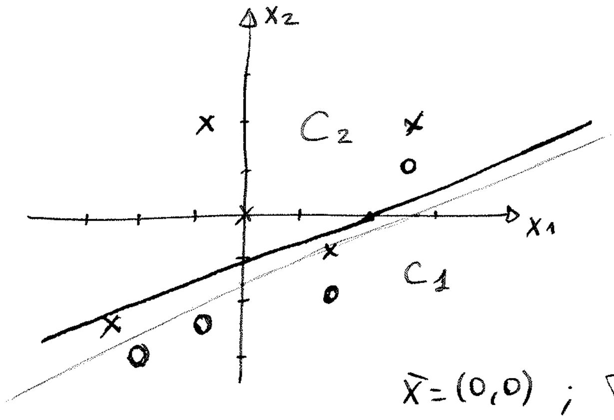
Algoritmo: reg_logistica_map (η, epocas, Datos D, σ^2)

- * Generamos $\bar{w}$ con valores $\in (-0'5 \;\; 0'5)$
- * para ie = 1 $\longrightarrow$ epocas
  - . Para $i = 1 \longrightarrow N$
    - Se calcula $\sigma_i = \sigma(\bar{w}^T \cdot \bar{x})$ // Prob de $\in C_1$
    - Se actualiza $\bar{w}$

      $$\bar{w} = \bar{w} - \eta (\sigma_i - t_i)\bar{x}_i - \frac{\eta}{N\sigma^2}\bar{w}$$

return.

Ejercicios



$w = (-8 \quad 4 \quad -8)$

4 o

5 x

$w_0 + w_1 x_1 + w_2 x_2 = 0$

$-8 + 4x - 8y = 0 \begin{cases} x = 2 \\ y = -1 \end{cases}$

¿Qué clase es cada parte de la recta? Miramos en el (0,0)

$\bar{x} = (0,0)$; $\sigma(-8 + 4 \cdot 0 - 8 \cdot 0) = \sigma(-8) = \dfrac{1}{1+e^8} < 0'5$

La prob. de ser de clase 1 es $< 0'5$, por lo que es de clase 2.

Asignar a o y x una clase (1 ó 2) minimizando el error.

$\left.\begin{array}{l} o \to C_1 \\ x \to C_2 \end{array}\right\}$ error $= \dfrac{3}{9}$ $\qquad$ $\left.\begin{array}{l} o \to C_2 \\ x \to C_1 \end{array}\right\}$ error $= \dfrac{6}{9}$

Asignación con error menor.

Actualizar w usando reg-logística-MAP para $\bar{x} = (-2'5, -2'5)$

$\eta = 0'5$, $\sigma^2 = 1$ y $N = 10$

$\bar{w} = \bar{w} - \eta(\sigma_i - t_i)\bar{x} - \dfrac{\eta}{N\sigma^2}\bar{w} = \begin{pmatrix} -8 \\ 4 \\ -8 \end{pmatrix} - 0'5(0'88 - 0)\begin{pmatrix} 1 \\ -2'5 \\ -2'5 \end{pmatrix} - 0'05 \begin{pmatrix} - \\ 4 \\ - \end{pmatrix}$

$= \begin{pmatrix} -8'4 \\ 3'78 \\ -6'5 \end{pmatrix}$

$\sigma_i = \sigma(\bar{w}^T \cdot \bar{x}) = \sigma(2) = \dfrac{1}{1+e^{-2}} = 0'88$

$\bar{w}^T \bar{x} = (-8 \quad 4 \quad -8)\begin{pmatrix} 1 \\ -2'5 \\ -2'5 \end{pmatrix} = -8 - 10 + 20 = 2$

recta : $-8'84 + 3'78 x_1 - 6'5 x_2 = 0$.

$x_1 = 0 \to x_2 = -\dfrac{8'84}{6'5}$

$x_2 = 0 \to x_1 = \dfrac{8'84}{3'78}$

* 2 clases con distr. normal en 4D

$$\Sigma_1 = \Sigma_2 = \mathbb{I} \qquad \mu_1 = (1, 1, 1, 1) \quad \mu_2 = (-1, -1, -1, -1)$$

¿Parámetros óptimos de la recta dada por Bayes?

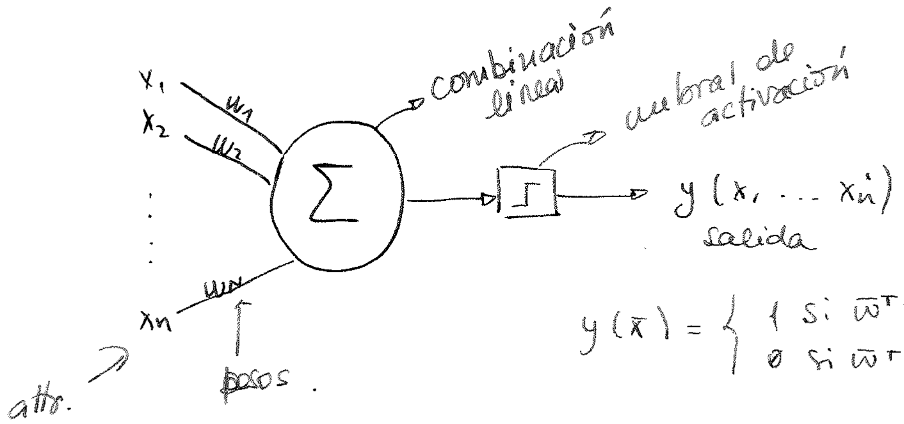
$$P(C_1) = P(C_2) \qquad \bar{w} = \bar{\mu}_1 - \bar{\mu}_2$$

$$w_0 = \frac{1}{2}\left(\bar{\mu}_2^2 - \bar{\mu}_1^2\right) + \ln \frac{P(C_1)}{P(C_2)}$$

$$\bar{w} = (1\ 1\ 1\ 1) - (-1\ -1\ -1\ -1) = (2\ 2\ 2\ 2)$$

$$w_0 = \frac{1}{2}\ (4-4) + \ln 1 = \emptyset$$

REDES NEURONALES



$$y(\bar{x}) = \begin{cases} 1 & \text{si } \bar{w}^T \bar{x} \geq \theta \\ 0 & \text{si } \bar{w}^T x < \theta \end{cases}$$

AND $\quad w_1 = w_2 = 1 \quad \theta = 1'5$

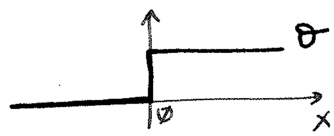
x_1	x_2	t	$x_1 w_1 + x_2 w_2$		$y(\bar{x})$
0	0	0	0	$0 > \theta \rightarrow$	0
0	1	0	1	$1 > \theta \rightarrow$	0
1	0	0	1	$1 > \theta \rightarrow$	0
1	1	1	2	$2 > \theta \rightarrow$	1

* Integración de umbral

$$y(x_1 \dots x_n) = \begin{cases} 1 & w_0x_0 + w_1x_1 + \dots + w_nx_n \geq 0 \\ 0 & \quad '' \qquad\qquad < 0 \end{cases}$$

$$\boxed{\begin{matrix} \theta = w_0 \\ x_0 = 1 \end{matrix}}$$

* Función escalon o de Heavinde



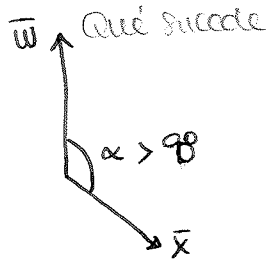
$$\theta(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

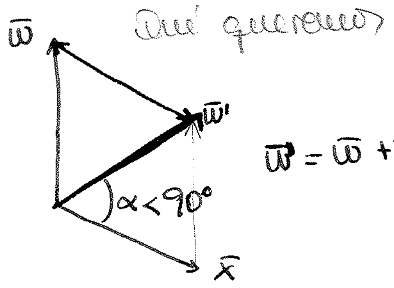
* Perceptron de Rosemblatt — PERCEPTRÓN SIMBLE —

$$y = (x_1 \dots x_n) = \theta(w_0x_0 + w_1x_1 + \dots + w_nx_n) = \theta(\bar{w}^T\bar{x})$$

* TIPOS DE ERROR para el perceptrón de Rosemblatt

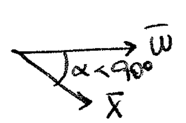
. si $y(\bar{x}) = 0$ y $\bar{x} \in C_1$ $(t=1)$

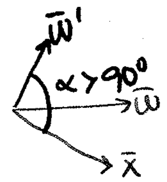




$\bar{w}' = \bar{w} + \bar{x}$

si $y(\bar{x}) = 1$ y $\bar{x} \in C_2$ $(t=0)$





$w' = \bar{w} - \bar{x}$

* ALGORITMO

. Generamos pesos aleatorios

. Para cada ejemplo $\bar{x}$

calculamos $y(\bar{x}) = \begin{cases} 1 & \text{si } \bar{w}^T\bar{x} \geq 0 \\ 0 & \text{si } \bar{w}^T\bar{x} < 0 \end{cases}$

. si $y(\bar{x}) = 0$ y $\bar{x} \in C_1 \longrightarrow \bar{w} = \bar{w} + \eta\bar{x}$

. si $y(\bar{x}) = 1$ y $\bar{x} \in C_2 \longrightarrow \bar{w} = \bar{w} - \eta\bar{x}$

* Criterio de Rosemblatt

delta

$$\underbrace{J(\bar{\omega})}_{\text{siempre positivo}} = \sum_{x \in D} \delta_x \bar{\omega}^T \bar{x}$$

donde $\delta_x = y(\bar{x}) - t(\bar{x})$

- si $y(\bar{x}) = \emptyset$ y $t = 1 \longrightarrow \delta_x = -1 \longrightarrow \delta_x \bar{\omega} \bar{x} \geq \emptyset$
- si $y(\bar{x}) = 1$ y $t = \emptyset \longrightarrow \delta_x = 1 \longrightarrow \delta_x \bar{\omega} \bar{x} >$

entonces minimizamos J con respecto a $\bar{\omega}$

$$\frac{\partial J(\bar{\omega})}{\partial \bar{\omega}} = \sum_{x \in D} \delta_x \bar{x} \Rightarrow \vec{\omega} = \bar{\omega} - \eta \delta_x \bar{x}$$

* ALGORITMO DE PERCEPTRÓN DE ROSEMBLATT $(\eta$, datos $d)$

$\bar{\omega}$ = pesos aleatorios $\in (-0'5, 0'5)$

mientras haya errores de clasificación

Para cada ejemplo $\bar{x} \in D$

es un valor no un vector

calculamos $\delta_x = y(\bar{x}) - t(\bar{x})$

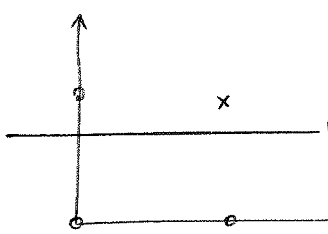
actualizamos pesos de $\bar{\omega}$: $\bar{\omega} = \bar{\omega} - \eta \delta_x \bar{x}$

* EJEMPLO (AND)

- Enunciado

$\eta = 0'25 \quad \bar{\omega} = (-0'3, 0, 0'4)$

x_0	x_1	x_2	t
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1



ej1 $\bar{x} = (1 \ 0 \ 0)$

$\delta_x = y(\bar{x}) - t(\bar{x}) = \emptyset$

$(-0'3, 0, 0'4) - (0, 0, 0) = (-0'3, 0, 0'4)$ La recta no cambia porque no hay error de clasificación

ej2 $\bar{x} = (1 \ 0 \ 1)$

$\delta_x = y(\bar{x}) - t(\bar{x}) = 1$

$\hookrightarrow \bar{w}^T \quad \begin{pmatrix} -0'3 \\ 0 \\ 0'4 \end{pmatrix} (1 \ 0 \ 1) = -0'3 + 0 + 0'4 = 0'1 \ \cancel{>} \emptyset$

$y(\bar{x}) = 1$

$\bar{w} = (-0'3, 0, 0'4) - 0'25 \cdot 1 \ (1, 0, 1) = (-0'55, 0, 0'15)$

ej3 $\bar{x} = (1, 1, \emptyset)$ la recta no cambia porque no hay error de clasificación.

ej4 $\bar{x} = (1, 1, 1)$

$\delta_x = y(\bar{x}) - t(\bar{x}) = -1$

$\hookrightarrow \begin{pmatrix} -0'55 \\ 0 \\ 0'15 \end{pmatrix} (1, 1, 1) = -0'55 + 0'15 < \emptyset$

$y(\bar{x}) = \emptyset$

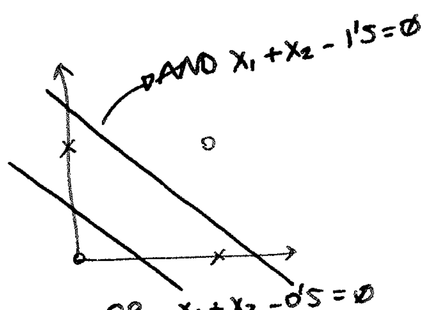
$\bar{w} = (-0'55, 0, 0'15) \underbrace{- 0'25(-1)}_{\text{positivo}} (1, 1, 1) = \underline{\underline{(-0'20, 0'25, 0'40)}}$

recta: $x_1 = 0 \rightarrow x_2 = 3/4$

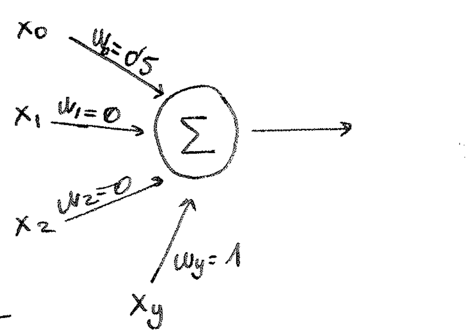
$x_2 = 0 \rightarrow x_1 = 1'1$

EJEMPLO XOR

$(x_1 - x_2)^2$	x_0	x_1	x_2	t
0	1	0	0	0
1	1	0	1	1
1	1	1	0	1
0	1	1	1	0



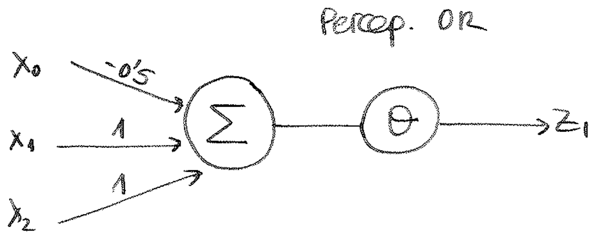
Necesitamos una capa para separar el problema



Estas rectas son las fronteras de los siguientes perceptrones:

$z_1 = \theta(x_1 + x_2 - 0'5)$

Percep. OR



x_1	x_2	t	z_1	z_2
0	0	0	0	0
^	1	1	1	0
		1	1	0
	1	0	1	1

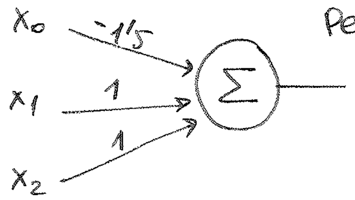
$z_2 = \theta(x_1 + x_2 - 1'5)$

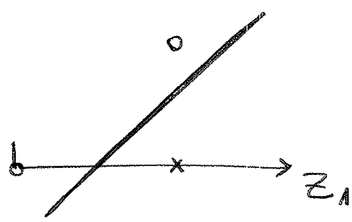
Recta por dos puntos

$(y - y_0) = m(x - x_0)$

$m = \dfrac{(y_2 - y_1)}{(x_2 - x_1)}$

Per





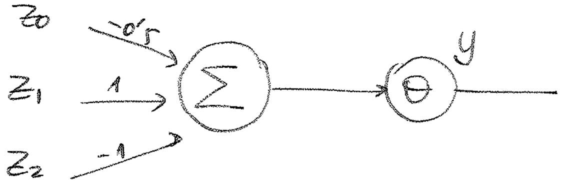
PESOS

$w_0 = -0'5$

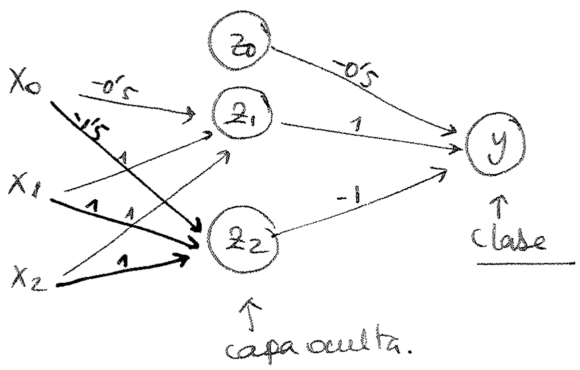
$w_1 = 1$

$w_2 = -1$

$\longrightarrow y = \theta(z_1 - z_2 - 0'5)$ Ahora el problema es separable linealmente.

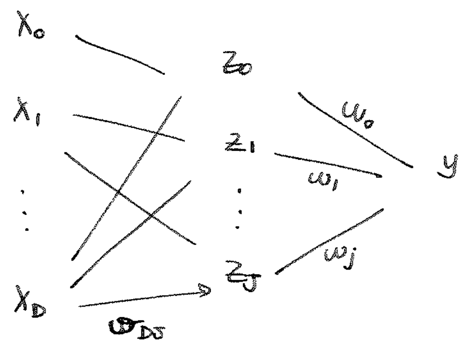


Todo esto da lugar al perceptron multicapa.



* Regla delta

- Sustituimos θ por σ, es decir, sigmoidal en vez de escalón
- Para perceptrones de una capa esto equivale a reg. logística.



D: dimensión del problema

J: nº neuronas en la capa oculta

v_{DJ}: peso entre neurona D de entrada y la neurona J (oculta)

w_j: peso entre neurona j (oculta) y salida.

Con una ~~neurona~~ σ en cada neurona

$$n^{\underline{o}} \text{ params} = (D+1) \times J + J + 1$$

$$\bar{x} = (1, x_1 \dots x_D) \quad x_0 = 1 \qquad \bar{z} = (1, z_1 \dots z_J) \quad z_0 = 1$$

Salida de la red:

$$y(\bar{x}) = \sigma(\bar{w}^T \bar{z}) = \sigma\left(\sum_{j=0}^{J} w_j z_j\right) = \sigma\left(\sum_{j=0}^{J} w_j \, \sigma\left(\sum_{d=0}^{D} x_d v_{dj}\right)\right)$$

$$z_j = \sigma\left(\sum_{d=0}^{D} x_d v_{dj}\right)$$

matriz $V = \begin{pmatrix} v_{01} & \dots & v_{0J} \\ \vdots & \ddots & \vdots \\ v_{D1} & \dots & v_{DJ} \end{pmatrix}$

$D = t_1 \dots t_N$
$I = \bar{x}_1 \dots \bar{x}_n$

$\begin{cases} t_1 \; 1 \text{ si } \bar{x}_i \in C_1 \\ t_2 \; 0 \text{ si } \bar{x}_i \in C_2 \end{cases}$

$$y(\bar{x}) = P(C_1 | \bar{x}, V, \bar{w})$$

* Resolvemos por ML

verosimilitud es

$$P(t_1 \dots t_n | V, \bar{w}, \bar{x}_1 \dots \bar{x}_n) = \prod_{n=1}^{N} \left[P(C_1 | \bar{x}_n, V, \bar{w})^{t_n} \cdot \left(1 - P(C_1 | \bar{x}_n, V, \bar{w})\right)^{1-t_n} \right] =$$

$$= \prod_{n=1}^{N} y(\bar{x}_n)^{t_n} (1 - y(\bar{x}_n))^{1-t_n}$$

Aplicamos $-\ln$

$$E(v, \bar{w}) = -\ln p(t_1 \dots t_N \mid \bar{x}_1 \dots \bar{x}_N, v, \bar{w}) = \sum_{n=0}^{N} E_n(v, \bar{w})$$

$$E_n(v, \bar{w}) = -t_n \ln y(\bar{x}_n) - (1 - t_n) \ln (1 - y(\bar{x}_n))$$

Derivada comodín

$$\frac{\partial E_n}{\partial \beta} = -t_n \frac{1}{y(\bar{x}_n)} \cdot \frac{\partial y(\bar{x}_n)}{\partial \beta} + (1 - t_n) \frac{1}{1 - y(\bar{x}_n)} \frac{\partial y(x_n)}{\partial \beta} =$$

$$= -t_n (1 - y(\bar{x}_n)) \frac{\partial \bar{w}^T \bar{z}}{\partial \beta} + (1 - t_n) y(\bar{x}_n) \frac{\partial \bar{w}^T \bar{z}}{\partial \beta} =$$

$$= (y(\bar{x}_n) - t_n) \frac{\partial \bar{w}^T \bar{z}}{\partial \beta}$$

$$\frac{\partial y(\bar{x}_n)}{\partial \beta} = \frac{\partial}{\partial \beta} \frac{1}{1 + e^{-\bar{w}^T \cdot \bar{z}}} = y(\bar{x}_n)(1 - y(\bar{x}_n)) \frac{\partial \bar{w}^T \bar{z}}{\partial \beta}$$

Optimizamos respecto ~~de~~ a w_i $(\beta \to w_i)$

$$\frac{\partial \bar{w}^T \bar{z}}{\partial w_i} = z_i \overset{\text{desarrollo}}{=} \frac{\partial}{\partial w_i} (w_0 z_0 + w_i \cdot z_i + \dots + w_J z_J) = z_j$$

$$\rightarrow \frac{\partial E_n}{\partial w_i} = (y(\bar{x}) - t_n) z_i$$

Optimizamos respecto a v_{pq} $(\beta \longrightarrow v_{pq})$

$$\frac{\partial \bar{w}^T \bar{z}}{\partial v_{pq}} = \frac{\partial}{\partial v_{pq}} \sum_{j=0}^{J} \bar{w}_j \bar{z}_j = \sum_{j=0}^{J} w_j \frac{\partial}{\partial v_{pq}} \overbrace{\sigma\left(\sum_{d=0}^{D} x_d v_{pq}\right)}^{z_j} =$$

$$= \sum_{j=0}^{J} w_j z_j (1 - z_j) \sum_{d=0}^{D} x_d \frac{\partial v_d}{\partial v_{pq}} = \sum_{j=0}^{J} w_j z_j (1 - z_j) \underbrace{\sum_{d=0}^{D} \delta_{pp} \delta_{jq}}_{x_p} =$$

$$w_q z_q (1 - z_q) x_p$$

$$d_{ij} = \begin{cases} 1 & \text{si } i = j \\ \emptyset & \text{si } i \neq j \end{cases}$$

$$\left[\frac{\partial E_u}{\partial v_{pq}} = \left(y(\bar{x}_u) - t_u \right) \frac{\partial \bar{\omega}^T \bar{z}}{\partial v_{pq}} = \left(y(\bar{x}) - t_u \right) w_q z_q (1 - z_q) x_p \right]$$

* Reglas de actualización

$$w_1 \longleftarrow w_i - \eta \left(y(\bar{x}_n) - t_n \right) z_i$$

$$v_{pq} \longleftarrow v_{pq} - \eta \left(y(\bar{x}_n) - t_n \right) w_q z_q (1 - z_q)\, x_p$$

ALGORITMO BATCH PROPAGATION – 2 CLASES (η, epocas, D_{train})

inicializamos $v, \bar{w}$ a valores ente $[-0'5, 0'5]$

for $ie = 1 \longrightarrow$ epocas

for $n = 1 \longrightarrow N$

$$z_q = \sigma\left(\sum_{d=0}^{D} v_{dq} \cdot \bar{x}_d \right) \quad // \; q = 1 \longrightarrow J$$

$$y(\bar{x}_n) = \sigma(\bar{w}^T \cdot \bar{z})$$

$$\delta_n = y(\bar{x}_n) - t_n$$

$$\Delta_q = \delta_n w_q z_q (1 - z_q) \quad // \; q = 1 \longrightarrow J$$

Actualización $\longrightarrow$ $w_q = w_q - \eta \delta_n z_q \quad // \; q = 1 \longrightarrow J$

$\longrightarrow$ $v_{pq} = v_{pq} - \eta \Delta_q x_p \quad // \; p = 1 \longrightarrow D$

MLP para problemas multiclase (varias salidas y)

K = nº clases

J = nº de neuronas en capa oculta

D = nº attr.

v_{dj} = peso entre neurona d (entrada) y j (oculta)

w_{jk} = peso entre neurona j (oculta) y k (salida)

$y_k(\bar{x})$ = prob. de que $\bar{x} \in C_k$

$$y_k = p(C_k \mid V, W, \bar{x})$$

por teorema de Bayes

$$P(C_k|\bar{x}) = \frac{P(\bar{x}|C_k)P(C_k)}{\sum_{k=1}^{K} P(\bar{x}|C_k)P(C_k)} = \frac{\exp(a_k)}{\sum_{k=1}^{K}\exp(a_k)}$$

$$a_k = \ln P(\bar{x}|C_k)P(C_k)$$

Ahora suponemos que a_k es la integración y_k de las salidas de la capa oculta, esto es:

$$a_k = \sum_{j=1}^{J} z_j w_{jk} \qquad y_k = P(C_k|V,W,\bar{x}) = \frac{\exp(a_k)}{\sum_{k=1}^{K}\exp(a_k)} =$$

$$= \frac{\exp\left(\sum_{j=1}^{J} w_{jk} z_j\right)}{\sum_{k=1}^{K}\exp\left(\sum_{j=1}^{J} w_{jk} z_j\right)}$$

Consideramos los datos

$D_{train} = (\{\bar{x}_n\ E_n\}\ n=1\dots N)$. Ahora E_n es un vector de todo $\emptyset$ excepto en la posición de la clase

$$\text{si } \bar{x} \in C_k \longrightarrow \begin{cases} t_k = 1 \\ t_j = \emptyset \quad \forall_j \neq k \end{cases}$$

$H = V, W$

$D = E_1 \dots E_N$

$I = \bar{x}_1 \dots \bar{x}_n$

Maximizado por MV (ML)

$$P(D|H,I) = P(E_1 \dots E_n | V,W,\bar{x}_1 \dots \bar{x}_n) =$$

$$= \prod_{n=1}^{N} P(E_n|V,W,\bar{x}_n) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

$$P(E_n|V,W,x_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

... Aplicamos $-\ln$

$$E(V,W) = -\ln P(E_1 \dots E_N | V,W,\bar{x}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

... calculamos un término

$$E_n(V,W) = \sum_{k=1} t_{nk} \ln y_{nk}$$

Derivada comodín

$$\frac{\partial E_n}{\partial \beta} = -\sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial \beta} = -\sum_{k=1} t_{nk} \left(\frac{\partial a_{nk}}{\partial \beta} - \sum y_{nk} \frac{\partial a_{nr}}{\beta} \right)$$

derivada softmax $\qquad \beta \equiv w_{qr}$

$$\frac{\partial y_{nk}}{\partial \beta} = y_{nk} \left(\frac{\partial a_{nk}}{\partial \beta} - \sum_{r} y_{nr} \frac{\partial a_{nr}}{\partial \beta} \right)$$

$$\frac{\partial E_n}{\partial w_{qr}} = \sum^{K} (y_{nk} - t_{nk}) \frac{\partial \sum_{j=1}^{K} z_j w_{jk}}{\partial w_{qr}} =$$

$$= \sum_{k} \sum_{j} (y_{nk} - t_{nk}) z_j \delta_{jq} \delta_{kr} = (y_{nk} - t_{nr}) z_q$$

$\beta = v_{pq}$

$$\frac{\partial E_n}{\partial v_{pq}} = \sum_{k}^{K} (y_{nk} - t_{nk}) \frac{\partial a_{nk}}{\partial v_{pq}} = \sum (y_{nk} - t_{nk}) \sum_{j} w_{jk} \cdot z_j \cdot$$

$$\cdot (1 - z_j) \sum_{i} x_n \delta_{ip} \delta_{jq} =$$

$$= \sum_{k=1}^{K} (y_{nk} - t_{nk}) w_{qk} z_{nq} (1 - z_{nq}) \bar{x}_{np}$$

Bach propagation nc (η, epocas, Dtrain)

Inicializamos v y w a valores entre [-0'5, 0'5]

for ie = 1 $\longrightarrow$ EPOCAS.

for ~~n~~ = 1 $\longrightarrow$ N

Calculamos la red:

$$z_j = \sigma\left(\sum_i x_{ni} v_{ij}\right) \quad // J * D$$

$$a_k = \sum_{j=0}^{J} z_j w_{jk} \quad // D * K$$

$$y_k = \exp(a_k) / \sum_{k=1}^{k} \exp(a_k) \quad // K$$

Para cada ud. de salida $1 \to k$

$$\Delta_k = y_k - t_{nk}$$

Para cada neurona oculta $1 \to J$

$$\Delta_j = \sum_{k=1} \Delta_k w_{jk} z_j (1 - z_j)$$

actualizamos

$$w_{jk} = w_{jk} - \eta \Delta_k z_j$$

$$v_{ij} = v_{ij} - \eta \Delta_j \bar{x}_{ni}$$

ret u, w

ALGORITMOS GENÉTICOS

Darwin

- Población de individuos compartiendo recursos.
- ∃ mecanismos de adaptación al entorno
  - Reproducción
  - Mutación
- Individuos que mejor se adaptan sobreviven.

AG utiliza estas ideas para resolver problemas de optimización

- Problema (entorno)
- Función de fitness (mide cómo se adapta un individuo)
- Conjunto de soluciones (Población)
- Individuos codifican una solución
- Individuos con mayor fitness mayor prob. de reproducción.
- Los individuos se recombinan heredando parte del material genético de sus progenitores
- Hay mutaciones aleatorias.

<u>Antes de empezar</u> hay que concretar:

- Codificación de la solución en una cadena (cromosoma
- Tamaño población.
- Definir operadores de la evolución:
  - Selección progenitores
  - Recombinación
  - Mutación
  - Selección supervivientes
- Condición terminación

ag (fitness f, params de evolución) {

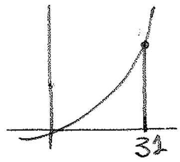
P = crear_población_aleatoria (tamaño)
while (no se cumpla criterio de terminación)

p' = selección_progenitores (P)
p' = recombinación (P')
p' = mutación (P')
P = selecciónSupervivientes (P, P')

end
return best (P)

}

Ej: encontrar el valor máximo de $f(x) = x^2$ en [0, 31] con x entero.
$\underline{}$ fitness



Cromosoma: Cadenas de 5 bits.

b_1	b_2	b_3	b_4	b_5

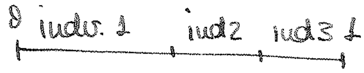
~~b está entre 0 y~~
b puede valer 0 ó 1.

$$x = \sum_{i=1}^{5} b_i \, 2^{5-i}$$

Tamaño población: T=4

. Selección progenitores: aleatoria y proporcional a fitness.
(ruleta con pesos):
↳ Ej

	f
individuo 1	100
indiv. 2	25
individ. 3	50



. Recombinación: Cruce en un pto.

0	1	0	0	1
1	1	0	0	0

⇨

0	1	0	0	0
1	1	0	0	1

. Mutación: cambio de bit con prob. P_m (pequeña)
ejm: $P_m = 1/20$
↳ T=4 * b=5 = 20

. Selección de supervivientes : Sustitución

PROB. INICIAL	X	FITNESS	$f_i / \sum f_i$	nº esperado en la sig iteración	num. select
01101	13	169	0'14	0'58	1
11000	24	576	0'49	1'97	2
01000	8	64	0'06	0'22	0
10011	19	361	0'31	1'23	1

$\bar{f} = 293.$

SELECCIÓN PROG	P.CRUCE	HIJOS	X	fi	MUTACIÓN	X	fi
0110|1	4	01100	12	144	01100		
1100|0		11001	25	625	11001		
11|000	2	11011	27	729	11011		
10|011		10000	16	256	10$\boxed{1}$00	20	400

$\bar{f} = 439$

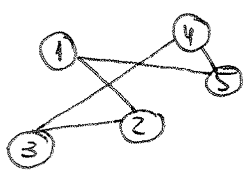
↳ mutación

Prob del viajante

5 ciudades

3	4	5	1	2

permutación



fitness = 1/distancia

recombinación

3	4	5	1	2
2	4	3	5	1

3	4	5	2	1
2	4	3	5	1

En el orden en el que aparecen (abajo)

(arriba)

mutación: Cambiar 2 nºs.

2	4	3	5	1

⇒

2	1	3	5	4

• Optimización pesos del perceptrón multicapa

pesos

v_{01}	v_{02}	...	v_{0j}	w_{01}	w_{02}	...	w_{0k}

$$fit = \frac{1}{error + 0'005}$$

$$fit2 = precisión.$$

Recombinación: cruce uniforme, en cada posición se tira una moneda

Mutación: Añadir ruido gaussiano $\approx N(0,1)$

29 Noviembre

* TEORÍA DE ESQUEMAS

Un bloque o esquema es una cadena (para codificación binaria) compuesta de: 0, 1 y $\boxed{*}$ **no definido**

$H = 1 * * * 0$ Todas las cadenas empiezan por 1 y terminan por 0.

$O(H)$: orden del esquema H. No de bits definidos

$\delta(H)$ = longitud del esquema. Distancia entre dos bits definidos más lejanos.

ej: $H = * 1 0 * 0$ $\delta(H) = 3$.

Para cadenas de longitud l, tendremos 2^l cromosomas posibles; $(2+1)^l$ esquemas.

Un cromosoma es instancia de 2^l esquemas

ej: $1 0 1 \Rightarrow$ está en 2^3 esquemas

101	10*	*01	***
	1*1	*0*	
	1**	**1	

Una población de N cromosomas representa $O(N2^l)$ esquem.

Inicialmente tendremos :

$N/2 \quad 1 * * * \dots$

$N/2 \quad \emptyset * * * \dots$

$n_H(t)$: nº de instancias de H en tiempo t.

$f_i(t)$: fitness individuo i en tiempo t.

$\bar{f}(t)$: fitness medio población en tiempo t.

$\bar{f}_H(t)$: fitness medio de las instancias que cumplen esquema H en tiempo t.

$\langle n_H(t+1) \rangle$ = nº esperado en tiempo t+1 de instancias de H.

Selección de progenitores proporcional a fitness

fitness : $\dfrac{f_i(t)}{\sum_{i=1}^{N} f_i(t)}$

$$\langle n_i(t+1) \rangle = N \, f_i(t) \Big/ \sum f_i(t) = f_i(t) \Big/ \underbrace{\bar{f}(t)}_{\text{media}}$$

$$\langle n_H(t+1) = N \sum_{i \in H} f_i(t) \Big/ \sum_{i=1}^{N} f_i(t) = \frac{n_H(t) \bar{f}_H(t)}{\bar{f}(t)}$$

$$\bar{f}_H(t) = \frac{1}{n_H(t)} \sum_{i \in H} f_i(t)$$

Si $\bar{f}_H(t) > f(t) \Rightarrow$ ~~Es~~ más probable que $n_H(t+1) > n_H(t)$

Añadimos cruce (cruce en un punto)

ej: $0|1110$ → 11110
$1|0101$ → 00101

$\boxed{*1**0}$

$$\boxed{S_c(H) = 1 - \frac{\delta(H)}{l-1}}$$ Prob. de que sobreviva el esquema al cruce.

→ mutación

P_m = prob. de mutación

Pro mutación de un bit $= 1 - P_m$

P supervivencia al esquema $= (1 - P_m)^{O(H)}$ $O(H)$ → nº de bits activos o definidos en el esquema.

$$\langle n_H(t+1) \rangle \geq \frac{\bar{f}_H(t)}{\bar{f}(t)} n_H(t) \left(1 - P_c \frac{\delta(H)}{l-1}\right)(1 - P_m)^{O(H)}$$

P_c → prob. de cruce (1)

4 Dic 2013

Ejemplo

$f(x) = x^2$

$\text{fitness} = \sum_{i=1}^{5} (b_i 2^i)^2$

$H_1 = 1****$
$H_2 = *10**$
$H_3 = 1***0$

Prob. inicial	x	f
0 1 1 0 1	13	169
1 1 0 0 0	24	576
0 1 0 0 0	8	64
1 0 0 1 1	19	361

$\bar{f} = 293$

$P_m = 1/20$

$P_c = 1$

$H_1 = 1****$

$f_H(t) = \frac{576+361}{2}$

$\text{P.SELEC}(H_1) = \frac{576+361}{2} \cdot \frac{1}{293} \cdot 2 = 3'2$

$S_c(H_1) = \left(1 - \frac{\cancelto{0}{\delta(H_1)}}{l-1}\right) = 1$

$S_m(H_1) = (1 - P_m)^{O(H_1)} = 19/20$

$\langle n_H(t+1) \rangle \geq S_c(H_1) S_m(H_1) \ldots$

$H_2 = *10**$ $\qquad \bar{f}_{H_2} = \frac{576 + 64}{2}$

\# cuando los bits están en los extremos $S_c = \emptyset$

$S_s(H_2) = \frac{576+64}{2} \cdot \frac{1}{293} \cdot 2 = \underline{2'18}$

$S_c(H_2) = \left(1 - \frac{\delta(H_2)}{\ell - 1}\right) = \left(1 - \frac{1}{5-1}\right) = \underline{0'75}$

$S_M(H_2) = (1 - P_M)^{O(H_2)} = (19/20)^2$

$\rightarrow$ # de bits que vienen definidos.

$\langle n_H(t+1)\rangle \geq \underline{\underline{2'95}}$

$H_3 = 1***0$

$S_s(H_3) = \underline{1'96}$

$S_c(H_3) = \emptyset$ $\qquad \langle n_H(t+1)\rangle \geq \emptyset$

$S_M(H_3) = (19/20)^2$

Ejemplo 2 $\quad$ **Enunciado.**

cadena	fit
10001	20
11100	10
00011	5
01110	15

$P_m = 0'01$ $\qquad \bar{f} = 12'5$

$P_c = 0'9$

$H_1 = 1****$

$H_2 = 0**1*$

¿ $n_{H_1}(t+1)$ y $n_{H_2}(t+1)$?

H_1

$\bar{f}_{H_1}(t) = \frac{20+10}{2} = 15$

$S_s(H_1) = \frac{15}{12'5} \cdot 2 = 2'4$

$S_c(H_1) = (1 - 0'9)\frac{0}{4} = 1$

$S_M(H_1) = (1 - 0'01)^1 = 0'99$

$\langle n_{H_1}(t+1)\rangle \geq \underline{2'376}$

H_2

$f_{H_2}(t) = \frac{5+15}{2} = 10$

$S_s(H_1) = \frac{10}{12'5} \cdot 2 = 1'6$

$S_c(H_2) = 1 - 0'9\,\frac{3}{4} = 0'075$ $\qquad \langle n_{H_2}(t+1)\rangle \geq \underline{\underline{1'176}}$

$S_M(H_2) = 0'99^2 = 0'98$

CLUSTERING (agrupamiento)

Para agrupar necesitamos:

* ~~Atributos~~
* Medida de proximidad o similitud.
* Algoritmo de clústering.
* (Validación e interpretación de resultados)

Definición de Clustering

$X = \{ \{\bar{x}_i\} \; i = 1, \dots, N \}$ definimos m-clustering como una partición en m subconjuntos $C_1, C_2, \dots, C_m$ tal que ningún grupo esté vacío y $C_i \cap C_j = \emptyset \;\; \forall i \neq j$.

Medidas de proximidad

sean $\bar{x}$ y $\bar{x}' \in X$ la métrica de distancia d ha de cumplir:

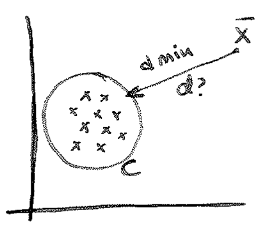
* $0 \leq d(\bar{x}, \bar{x}') \leq \infty$
* $d(\bar{x}, \bar{x}') = 0 \iff \bar{x} = \bar{x}'$
* $d(\bar{x}, \bar{x}') = d(\bar{x}', \bar{x})$
* $d(\bar{x}, \bar{x}'') \leq d(\bar{x}, \bar{x}') + d(\bar{x}', \bar{x}'')$

· Distancias entre puntos

Euclídea $d(\bar{x}, \bar{x}') = \sqrt{\sum_{d=1}^{D} (x_d - x'_d)^2}$

l_1 o Manhattan $d(\bar{x}, \bar{x}') = \sum_{d=1}^{D} | x_d - x'_d |$

· Distancia entre un punto y un cluster.



- Mínimo

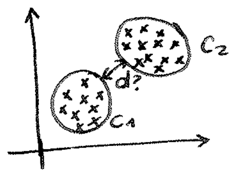
$d(\bar{x}, C) = \min_{x' \in C} (\bar{x}, \bar{x}')$

- Máximo

$d(\bar{x}, c) = \max_{x' \in C} (\bar{x}, \bar{x}')$

- Media

- **Distancia entre clusters**



- Mínimo (enlace simple)

$$d_{min}(C_1, C_2) = \min_{\bar{x} \in C_1, \bar{x}' \in C_2} d(\bar{x}, x')$$

- Máximo (enlace completo)

$$d_{max}(C_1, C_2) = \max_{\bar{x} \in C_1, \bar{x}' \in C_2} d(\bar{x}, \bar{x}')$$

- Media

$$d_{med}(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{\bar{x} \in C_1, \bar{x}' \in C_2} d(\bar{x}, \bar{x}')$$

- **Clustering jerárquico aglomerativo.**

Devuelve una secuencia de agrupamientos

$$R_0 \subset R_1 \subset R_2 \subset \dots \subset R_{N-1}$$

tal que $R_0 = \{C_i = \{\bar{x}_i\}, i = 1 \dots N\}$ // cada punto en un cluster.

y $R_{N-1} = \{C_1 = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n\}\}$ // todos los puntos en 1 cluster.

<u>ALGORITMO</u>

```
clus_jer_aglo (X = {x̄, ... x̄n}, métrica d)
    inicializamos R0 = {Ci = {xi}; i = 1, ..., N}
    n = ∅;
    do
        n = n+1;
        Buscar clusters Ci, Cj tal que d(Ci, Cj) sea mínima;
        Rn = {Rn-1 - Ci - Cj} ∪ {Ci ∪ Cj}
    hasta que sólo haya un cluster.
    Return R0 ... RN-1
```

Resulta útil definir una matriz de distancias D tal que D_{ij} sea $d(C_i, C_j)$

ej: $\bar{x}_1\ \bar{x}_2\ \bar{x}_3\ \bar{x}_4\ \bar{x}_5$

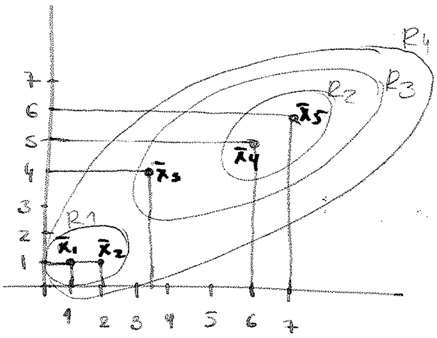
$\bar{X}_1 = (1,1)$

$\bar{X}_2 = (2,1)$

$\bar{X}_3 = (3'5,4)$

$\bar{X}_4 = (6,5)$

$\bar{X}_5 = (7,6)$



$$D = \begin{array}{c} \begin{matrix} \{\bar{x}_1\} & \{\bar{x}_2\} & \{\bar{x}_3\} & \{\bar{x}_4\} & \{\bar{x}_5\} \end{matrix} \\ \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 3'9 & 3'4 & 0 & & \\ 6'4 & 5'7 & 2'7 & 0 & \\ 7'8 & 7'1 & 4'6 & 1'4 & 0 \end{pmatrix} \end{array}$$

$R_0 = \{\{\bar{x}_1\}\{\bar{x}_2\}\{\bar{x}_3\}\{\bar{x}_4\}\{\bar{x}_5\}\}$

$d = d_{min}$ (enlace simple) con distancia euclidea.

$n=1$ $R_1 = \{\{\bar{x}_1, \bar{x}_2\}\{\bar{x}_3\}\{\bar{x}_4\}\{\bar{x}_5\}\}$

Actualizamos la matriz

$$D = \begin{array}{c} \begin{matrix} \{\bar{x}_1,\bar{x}_2\} & \{\bar{x}_3\} & \{\bar{x}_4\} & \{\bar{x}_5\} \end{matrix} \\ \begin{pmatrix} 0 & & & \\ 3'4 & 0 & & \\ 5'7 & 2'7 & 0 & \\ 7'1 & 4'0 & 1'4 & 0 \end{pmatrix} \end{array}$$

$n=2$ $R_2 = \{\{\bar{x}_1, \bar{x}_2\}\{\bar{x}_3\}\{\bar{x}_4, \bar{x}_5\}\}$

$$D = \begin{array}{c} \begin{matrix} \{\bar{x}_1,\bar{x}_2\} & \{\bar{x}_3\} & \{\bar{x}_4,\bar{x}_5\} \end{matrix} \\ \begin{pmatrix} 0 & & \\ 3'4 & 0 & \\ 5'7 & 2'7 & 0 \end{pmatrix} \end{array}$$

$n=3$ $R_3 = \{\{\bar{x}_1, \bar{x}_2\}\{\bar{x}_3, \bar{x}_4, \bar{x}_5\}\}$

$$D = \begin{array}{c} \begin{matrix} \{\bar{x}_1,\bar{x}_2\} & \{\bar{x}_3,\bar{x}_4,\bar{x}_5\} \end{matrix} \\ \begin{pmatrix} 0 & \\ 3'4 & 0 \end{pmatrix} \end{array}$$

$R_4 = \{\{\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5\}\}$

* Orden de ejecución

N puntos

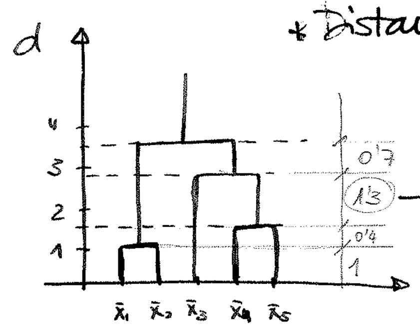
(N-1) pasos y en cada paso juntamos dos en el paso n

hay N-n grupos.

$$\underbrace{\frac{N(N-1)}{2} + \frac{(N-1)(N-2)}{2} + \dots + \frac{2\cdot 1}{2}}_{N-1 \text{ términos}} = O(N^3)$$

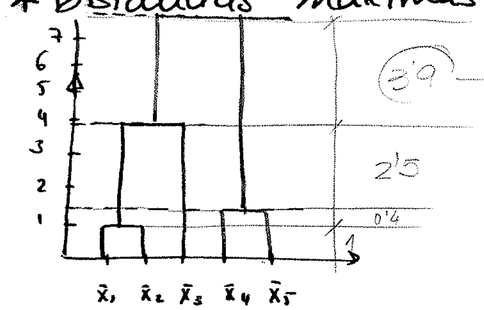
Dendograma

* Distancias mínimas



Partición más estable
$R_2 = \{\{\bar{x}_1, \bar{x}_2\}\{\bar{x}_3\}\{\bar{x}_4, \bar{x}_5\}\}$

* Distancias máximas



Partición más estab

$R_3 = \{\{\bar{x}_1, \bar{x}_2, \bar{x}_3\}\{\bar{x}_4, \bar{x}_5\}\}$

Clustering secuencial

Se procesan los puntos secuencialmente. Cada punto se asigna a un cluster o se crea un cluster nuevo.
Se usa distancia umbral.

clust_sec $(X = \{\bar{x}_1, \dots, \bar{x}_n\}$, métrica d, umbral θ) {

- $c_1 = \{\bar{x}_1\}$
- $m = 1$
- para cada patrón $\bar{x}_i$ con $i = 2 \dots N$
  - Busca el cluster C_k tal que $d(\bar{x}_i, C_k) = \min d(\bar{x}_i, C_j) \quad 1 \le j \le m$
  - si $d(\bar{x}_i, C_k) \le \theta$
    - $C_k = C_k \cup \{\bar{x}_i\}$
  - else
    - $m = m + 1$
    - $C_m = \{\bar{x}_i\}$

Ej: [$\{x_1, x_2, x_3, x_4, x_5\}$

$$D = \begin{pmatrix} 0 \\ 1 & 0 \\ 3'9 & 3'4 & 0 \\ 6'4 & 5'7 & 2'7 & 0 \\ 7'8 & 7'1 & 4'0 & 1'4 & 0 \end{pmatrix}$$

$\theta = 3$

$d = d_{min}$] enunciado

$m = 1$

$C_1 = \{\bar{x}_1\}$

$\bar{x}_2$

$d(\bar{x}_1 = C_1) = 1 \leq \theta$

$C_1 = \{\bar{x}_1, \bar{x}_2\}$

$\bar{x}_3$

$d(\bar{x}_3, C_1) = 3'4 > \theta$

$m = 2$

$C_1 = \{\bar{x}_1, \bar{x}_2\}$ $C_2 = \{\bar{x}_3\}$

$\bar{x}_4$

$d(\bar{x}_4, C_1) = 5'7$

$d(\bar{x}_1, C_2) = 27$

cogemos el mínimo y aplicamos el umbral

$2'7 \leq \theta$

$C_1 = \{\bar{x}_1, \bar{x}_2\}$

$C_2 = \{\bar{x}_3, \bar{x}_4\}$.

$\bar{x}_5$ $d(\bar{x}_5, C_1) = 7'1$

$d(\bar{x}_5, C_2) = 1'4 \leq \theta$ ← min

Solución

$C_1 = \{\bar{x}_1, \bar{x}_2\}$

$C_2 = \{\bar{x}_3, \bar{x}_4, \bar{x}_5\}$

$m = 2$

Ej: clustering languages

enunc. [solo con ESP, FRA, HUN, ING, ITA, NOR $\theta = 4$]

$$D = \begin{pmatrix} 0 \\ 2 & 0 \\ 10 & 10 & 0 \\ 6 & 6 & 9 & 0 \\ 1 & 1 & 10 & 6 & 0 \\ 6 & 6 & 8 & 2 & 6 & 0 \end{pmatrix}$$

$\boxed{1^o}$ $C_1 = \{ESP\}$

$m = 1$

$\boxed{2^o}$ $d(FRA, C_1) = 2 \leq \theta$

$C_1 = \{ESP, FRA\}$

$\boxed{3^o}$ $d(HUN, C_1) = 10 > \theta$

$C_1 = \{ESP, FRA\}$

$C_2 = \{HUN\}$

$\boxed{4^o}$ $d(ING, C_1) = 6 > \theta$

$d(ING, C_2) = 9$

$C_1 = \{ESP, FRA\}$

$C_2 = \{HUN\}$

$C_3 = \{ING\}$

$\boxed{5^o}$ $\underset{min}{d}(ITA) = 1 \leq \theta$ $\quad d(ITA, C_2) = 10$

$d(ITA, C_3) = 6$

$C_1 = \{ESP, FRA, ITA\}$ $\quad C_2 = \{HUN\}$ $\quad C_3 = \{ING\}$

$\boxed{6^o}$ $d(NOR, C_1) = 6$ $\quad d(NOR, C_2) = 8$ $\quad \underset{min}{\underline{d(NOR, C_3)}} = 2 \leq \theta$

<u>Solución</u> $C_1 = \{ESP, FRA, ITA\}$

$C_2 = \{HUN\}$

$\boxed{m = 3}$ $C_3 = \{ING, NOR\}$

Ej: Jerarquía con dmean