Prácticas de Aprendizaje Automático

Trabajo 2: Complejidad de H y Modelos Lineales

Pablo Mesejo y Jesús Giráldez
Universidad de Granada
Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

1. Ejercicio sobre la complejidad de H y el ruido

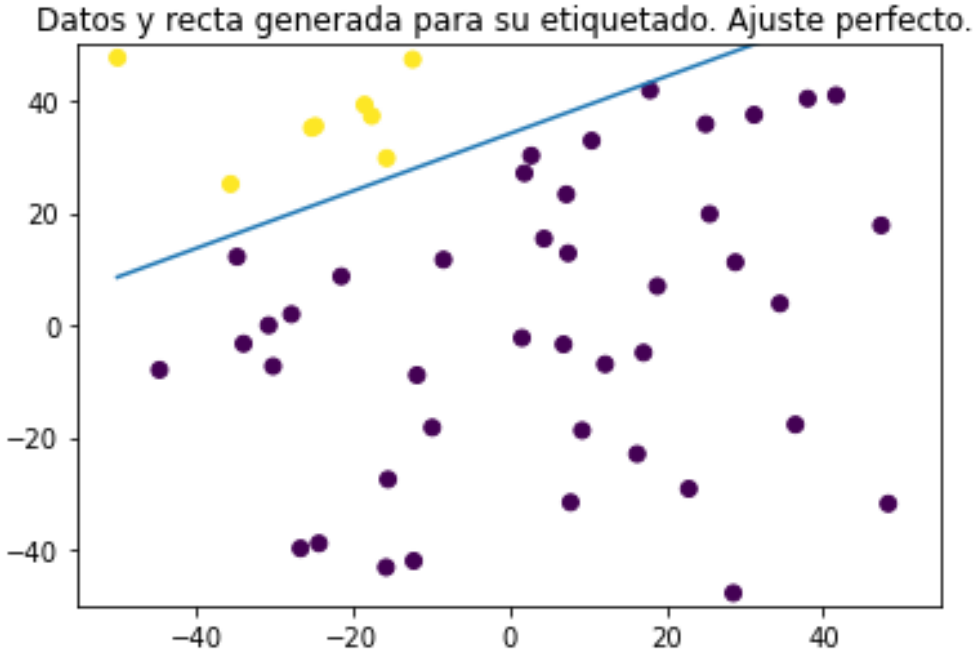
H es nuestro *Hypothesis Set*

Conjunto de fórmulas candidatas para resolver nuestro problema. Al aplicar un algoritmo de aprendizaje, somos capaces de quedarnos con una única hipótesis final (*g*, que intenta aproximar *f*, la función objetivo desconocida)

Hypothesis Set	Learning Algorithm
ANNs	Backpropagation
Perceptron	PLA
Linear Regression	Pseudoinverse

1. Ejercicio sobre la complejidad de H y el ruido

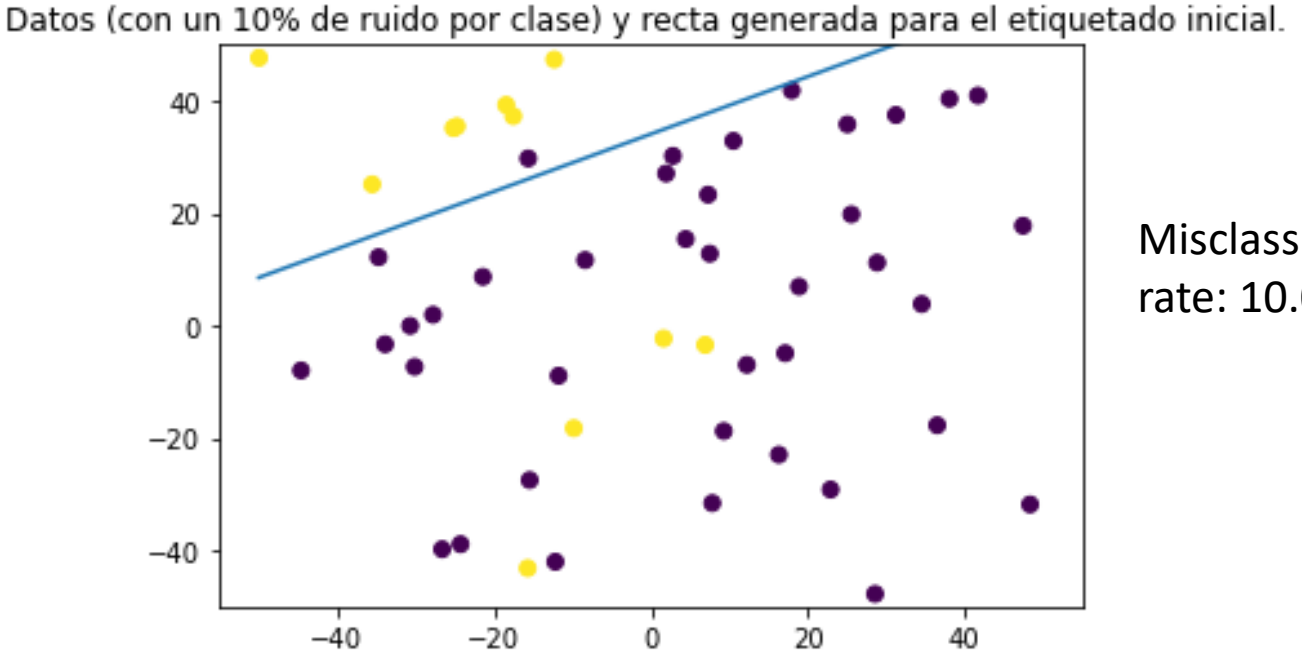
- Hay que generar una muestra de puntos 2D a la que vais a añadir una etiqueta usando el signo de la función **`f(x,y) = y-ax-b`**



Misclassification rate: 0.0%

1. Ejercicio sobre la complejidad de H y el ruido

- Modificar de forma aleatoria un 10% de las etiquetas positivas y otro 10% de las negativas



Misclassification rate: 10.0%

1. Ejercicio sobre la complejidad de H y el ruido

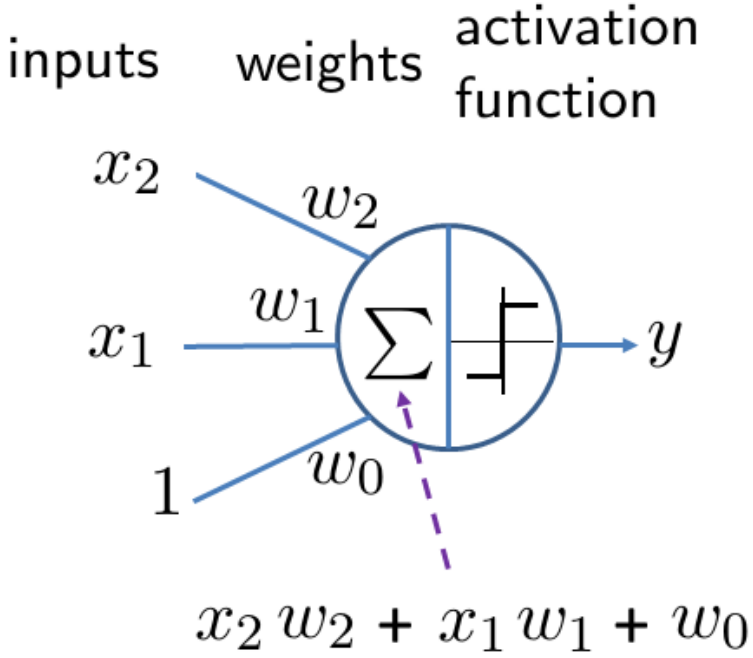
- Emplear otras funciones para definir la frontera de clasificación de los puntos de la muestra en lugar de una recta

$$f(x,y) = (x-10)^2 + (y-20)^2 - 400$$
$$f(x,y) = 0{,}5(x+10)^2 + (y-20)^2 - 400$$
$$f(x,y) = 0{,}5(x-10)^2 - (y+20)^2 - 400$$
$$f(x,y) = y - 20x^2 - 5x + 3$$

  - ¿Necesariamente funciones más complejas son mejores clasificadores (es decir, representan “mejores” bordes de decisión)?
  - ¿Es posible superar/mejorar ese 10% de error de clasificación?
  - ¿Qué pasa si repetimos el proceso con estas funciones más complejas (las empleamos para etiquetar los datos y luego metemos un 10% de ruido)? ¿Qué error de clasificación tenemos? ¿Es menor que ese 10%?

NOTA: se aconseja utilizar la función plot_datos_cuad proporcionada en la plantilla.

2. Modelos Lineales. Perceptrón

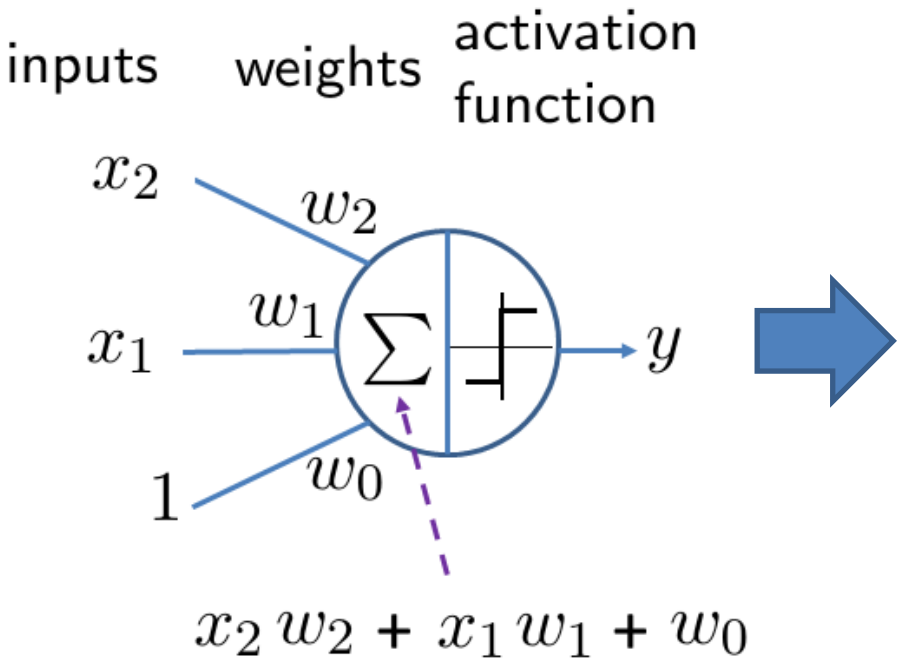


Referencias de apoyo sobre el perceptrón:

"Pattern Recognition and Machine Learning" (Christopher M. Bishop, 2006, págs. 192-196)

"Learning from Data" (Yaser S. Abu-Mostafa et al., 2012, págs. 5-8)

2. Modelos Lineales. Perceptrón



$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i\, x_i\right) + \quad w_0 \quad\right)$$

Introduce an artificial coordinate $x_0 = 1$:

$$h(\mathbf{x}) = \text{sign}\left(\sum_{i=0}^{d} w_i\, x_i\right)$$

In vector form, the perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x})$$

http://work.caltech.edu/slides/slides01.pdf slide 11

2. Perceptron Learning Algorithm (PLA)

- Given the data set $(\mathbf{x}_n, y_n), n = 1,2,\cdots,N$
- Step.1: Fix $\mathbf{w}_{\text{ini}}=0$
- Step.2: Iterate on the $\mathcal{D}$-samples improving the solution:
- repeat

  For each $\boldsymbol{x}_i \in \mathcal{D}$ **do**

  if: $sign(\boldsymbol{w}^T \boldsymbol{x}_i) \neq y_i$ **then**

  update $\mathbf{w}$: $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + y_i \mathbf{x}_i$

  else continue

  End for
- Until No changes in a full pass on $\mathcal{D}$

2. Perceptron Learning Algorithm (PLA)

¿ $E_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{sign(\boldsymbol{w}^T\boldsymbol{x}_i)\neq y_i}$ vs $E_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \max(0, -y_i\boldsymbol{w}^T\boldsymbol{x}_i)$?

- La *0-1 loss* (de la izquierda) no es derivable → problema a la hora de emplearla con gradiente descendente → se adapta dicha expresión a un formato más amigable para la optimización.

Now the gradient : $\nabla error = \frac{\partial \max(0, -y_n \mathbf{w}^T \mathbf{x}_n)}{\partial \mathbf{w}} = -y_n \mathbf{x}_n$ if $sign(y_n) \neq sign(\mathbf{w}^T \mathbf{x}_n)$
0 otherwise

- Ambas presentan un funcionamiento similar (dan 0 si la salida deseada y la obtenida coinciden, y dan otro valor si no es así).

2. Perceptron Learning Algorithm (PLA)

¿ $E_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{sign(\boldsymbol{w}^T\boldsymbol{x}_i)\neq y_i}$ vs $E_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \max(0, -y_i\boldsymbol{w}^T\boldsymbol{x}_i)$?

i) $y_n = -1 \;\&\; w^T x_n > 0$ → error → max(0, -(-1*valor_positivo)) = max(0, valor_positivo) = valor_positivo

ii) $y_n = +1 \;\&\; w^T x_n > 0$ → acierto → max(0, -(1*valor_positivo)) = max(0, valor_negativo) = 0

iii) $y_n = -1 \;\&\; w^T x_n < 0$ → acierto → max(0, -(-1*valor_negativo)) = max(0, valor_negativo) = 0

iv) $y_n = +1 \;\&\; w^T x_n < 0$ → error → max(0, -(1*valor_negativo)) = max(0, valor_positivo) = valor_positivo

2. Perceptrón. Intuición.

$$\boldsymbol{w_{new}} = \boldsymbol{w_{old}} + \boldsymbol{y_i} \cdot \boldsymbol{x_i}$$

- Definición algebraica de producto escalar:

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_1 \cdot b_1 + a_2 \cdot b_2 + a_3 \cdot b_3 + \ldots + a_n \cdot b_n$$

- Interpretación geométrica:

$\boldsymbol{a} \cdot \boldsymbol{b} = ||\boldsymbol{a}|| \cdot ||\boldsymbol{b}|| \cdot \cos(\vartheta)$, siendo ϑ el ángulo que forman $\boldsymbol{a}$ y $\boldsymbol{b}$, y siendo $||\boldsymbol{a}||$ y $||\boldsymbol{b}||$ la magnitud de ambos vectores (es decir, $||\boldsymbol{a}|| = \sqrt{{a_1}^2 + {a_2}^2 + \ldots + {a_n}^2}$)

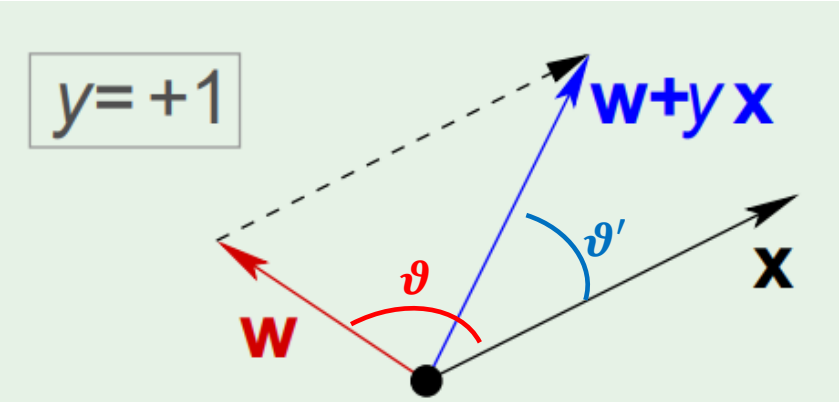
Se trata de un problema de clasificación binaria → solo nos interesa el signo de la anterior operación. Y, ¿de qué va a depender el signo? Del coseno.

Si $0° \leq \vartheta \leq 90°$→ $\cos(\vartheta) \geq 0$

Si $90° < \vartheta \leq 180°$→ $\cos(\vartheta) < 0$

2. Perceptron. Intuición.

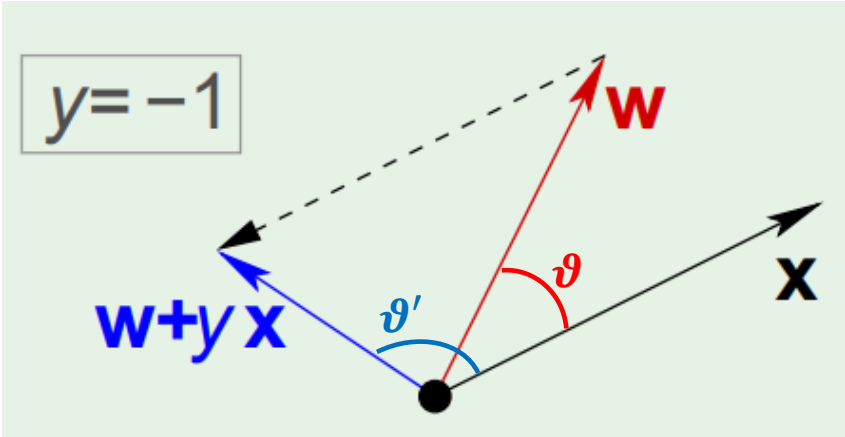
Caso 1: la salida deseada es positiva ($y_i = +1$), pero nuestra salida obtenida es negativa ($\boldsymbol{w} \cdot \boldsymbol{x} < 0$). Esto está asociado con un ϑ (ángulo entre w y x) $> 90°$.



¿Cómo solucionarlo? Vamos a sumar un vector a w, de modo que ϑ disminuya. Buscamos que $(\boldsymbol{w} + y \cdot \boldsymbol{x}) \cdot \boldsymbol{x} > 0$. De cumplirse esta última condición, ya estaríamos clasificando correctamente ese ejemplo.

2. Perceptrón. Intuición.

Caso 2: la salida deseada es negativa ($y_i = -1$), pero nuestra salida obtenida es positiva ($\boldsymbol{w} \cdot \boldsymbol{x} > 0$). Esto está asociado con un ϑ (ángulo entre w y x) $< 90°$.



¿Cómo solucionarlo? Vamos a restar un vector a w, de modo que ϑ aumente. Como $y = -1$, podemos realizar exactamente la misma operación de antes, buscando que $(\boldsymbol{w} + y \cdot \boldsymbol{x}) \cdot \boldsymbol{x} < 0$. De cumplirse esta última condición, ya estaríamos clasificando correctamente ese ejemplo.

2. Perceptrón. Intuición.

Referencias de interés:

- Abu-Mostafa ("**Lecture 01 - The Learning Problem**"): https://www.youtube.com/watch?v=mbyG85GZ0PI&t=1416s y https://home.work.caltech.edu/slides/slides01.pdf (diapositiva 12)

- University of Utah (The perceptron algorithm): https://www.cs.utah.edu/~zhe/pdf/lec-10-perceptron-upload.pdf

2. Perceptrón. Apuntes finales.

- Si **datos linealmente separables** → Convergencia garantizada
  - De lo contrario → No convergerá

- Es un algoritmo que **suele requerir muchas iteraciones para converger** (en problemas linealmente separables)
  - En nuestro caso, del orden de miles o decenas de miles de iteraciones…
  - Una iteración se corresponde con la presentación de un ejemplo

- **Dependiente de la inicialización y el número de iteraciones**

2. Modelos Lineales. Logistic Regression

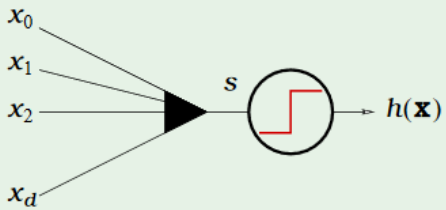
A third linear model

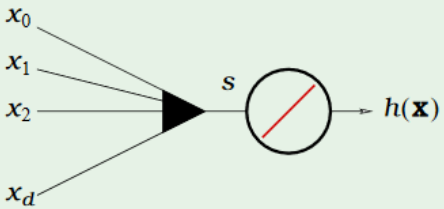
lineales en los parámetros del modelo

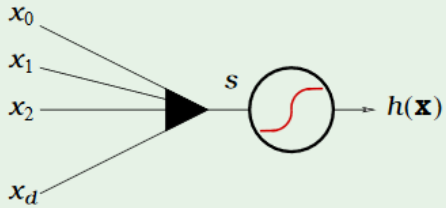
$$s = \sum_{i=0}^{d} w_i x_i$$

¡Ojo! Aunque usa el término "regresión" es una técnica de clasificación

linear classification	linear regression	logistic regression
$h(\mathbf{x}) = \text{sign}(s)$	$h(\mathbf{x}) = s$	$h(\mathbf{x}) = \theta(s)$







http://work.caltech.edu/slides/slides09.pdf slide 10

2. Modelos Lineales. Logistic Regression

The LR classifier is defined as

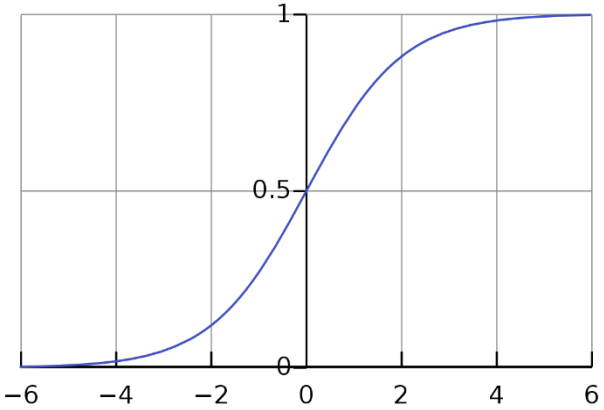
$$\sigma\left(f(\mathbf{x}_i)\right) \begin{cases} \geq 0.5 & y_i = +1 \\ < 0.5 & y_i = -1 \end{cases}$$

where $\sigma(f(\mathbf{x})) = \frac{1}{1+e^{-f(\mathbf{x})}}$

la salida está acotada entre 0 y 1 y, por tanto, se interpreta en términos probabilísticos

La función logística es un tipo de función sigmoide (como podría ser también la hiperbólica tangente)

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



2. Modelos Lineales. Logistic Regression

Logistic regression algorithm

Ajustad cuidadosamente el learning rate y N

1: Initialize the weights at $t = 0$ to $\mathbf{w}(0)$

2: **for** $t = 0, 1, 2, \ldots$ **do**

3: Compute the gradient

$$\nabla E_{\text{in}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^{\intercal}(t)\mathbf{x}_n}}$$

4: Update the weights: $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla E_{\text{in}}$

5: Iterate to the next step until it is time to stop

6: Return the final weights $\mathbf{w}$

inicializar los pesos a cero implica inicializar la probabilidad a 0.5

N: batch size

Parar el algoritmo cuando $||\mathbf{w}^{(t-1)} - \mathbf{w}^{(t)}|| < 0{,}01$, donde $\mathbf{w}^{(t)}$ denota el vector de pesos al final de la época t.

2. Modelos Lineales. Logistic Regression

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-y_n \mathbf{w}^{\mathsf{T}}\mathbf{x}_n}\right)$$

"cross-entropy" error

Para comprender mejor de dónde proviene esta expresión, se recomienda consultar https://home.work.caltech.edu/slides/slides09.pdf y https://www.youtube.com/watch?v=qSTHZvN8hzs

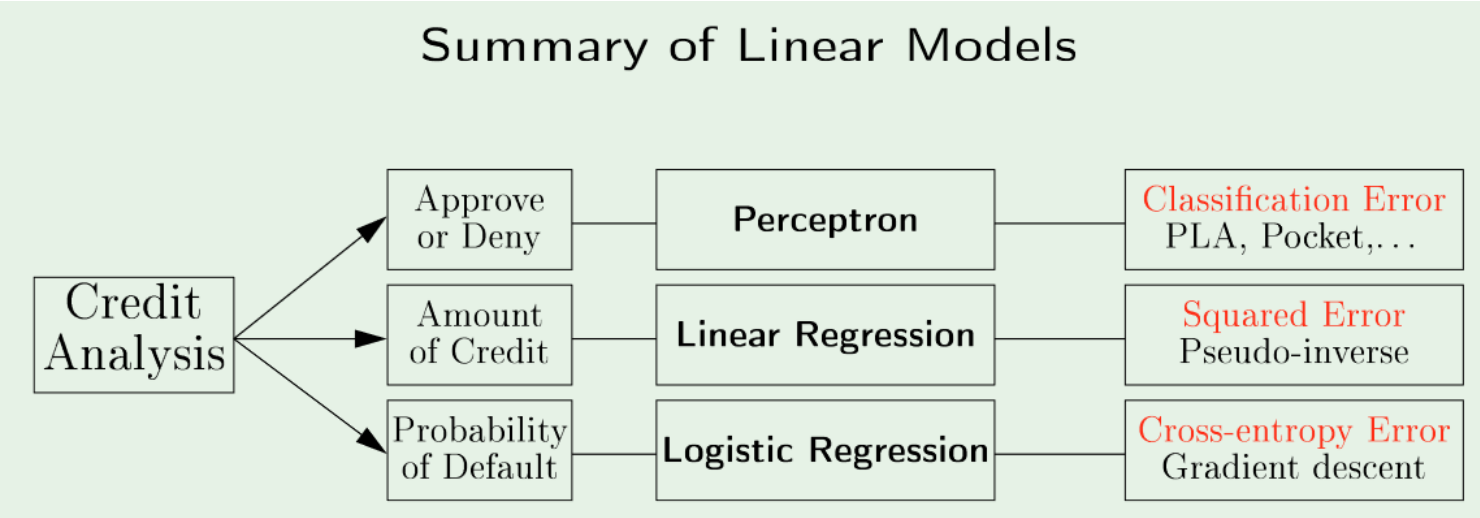
Comparativamente, recordad que en regresión lineal era

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{w}^{\mathsf{T}}\mathbf{x}_n - y_n)^2$$

Error cuadrático medio
(y este error se podía minimizar en *closed-form* con la pseudoinversa)

2. Modelos Lineales. Logistic Regression

Contextualizando todo un poco…



https://home.work.caltech.edu/slides/slides09.pdf

Más información sobre Logistic Regression (Andrew Ng; utiliza una notación diferente):
https://www.youtube.com/playlist?list=PLNeKWBMsAzboR8vvhnlanxCNr2V7ITuxy

2. Modelos Lineales. Logistic Regression

Por hacer las cosas un poco más concretas. Veamos una posible iteración con SGD:

$$\nabla E_{\text{in}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n \mathbf{x}_n}{1+e^{y_n \mathbf{w}^{\mathsf{T}}(t)\mathbf{x}_n}}$$

```
N = 5 #nuestro minibatch tiene 5 ejemplos/filas
η = 0.1 #learning rate
x = [[1.      1.48780677 1.2734803 ]
     [1.      0.23127434 0.8996494 ]
     [1.      1.29469747 0.61545333]
     [1.      0.64125328 0.45409605]
     [1.      0.20640655 0.75364972]]
       x_n0       x_n1        x_n2
y = [-1. 1. -1. -1. 1.]
w = [0.0 0.0 0.0]
      w_0  w_1  w_2
```

prod. elemento a elemento

$y_n \cdot x_n$ =
```
[[-1.      -1.48780677  -1.2734803 ]
 [ 1.       0.23127434   0.8996494 ]
 [-1.      -1.29469747  -0.61545333]
 [-1.      -0.64125328  -0.45409605]
 [ 1.       0.20640655   0.75364972]]
```

dot product

$y_n w^T x_n$ = `[0. 0. 0. 0. 0.]`

$e^{y_n w^T x_n}$ = `[1. 1. 1. 1. 1.]`

$1 + e^{y_n w^T x_n}$ = `[2. 2. 2. 2. 2.]`

$\frac{y_n \cdot x_n}{1+e^{y_n w^T x_n}}$ =
```
[[-0.5        -0.74390338 -0.63674015]
 [ 0.5         0.11563717  0.4498247 ]
 [-0.5        -0.64734873 -0.30772667]
 [-0.5        -0.32062664 -0.22704803]
 [ 0.5         0.10320327  0.37682486]]
```

$\nabla Ein = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n \cdot x_n}{1+e^{y_n w^T x_n}}$ = `[0.1 0.29860766 0.06897306]`

$w = w - \eta \nabla E_{in}$ = `[-0.01 -0.02986077 -0.00689731]`

`new w_0  new w_1  new w_2`

2. Modelos Lineales. Más ideas.

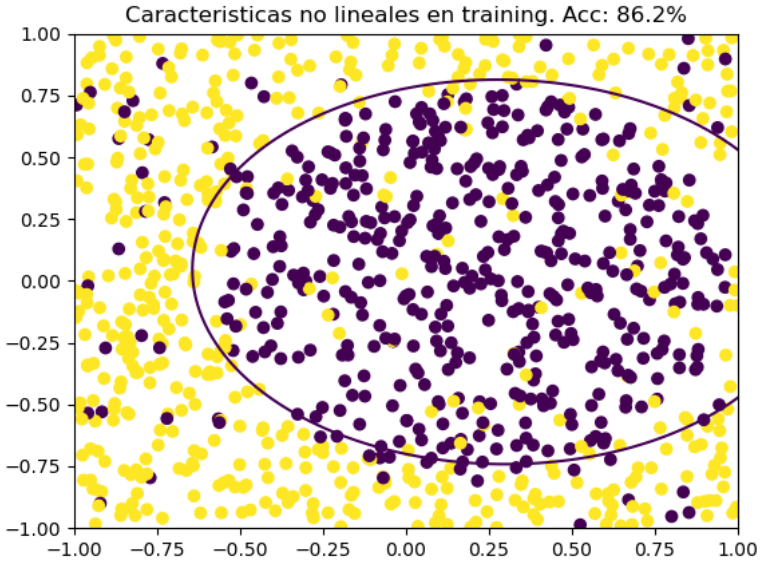
- Hablamos siempre de modelos lineales. Pero, ¿lineales en qué?
  - ¡En los pesos!
  - $w_0 + w_1x_1 + w_2x_2^2 = y$ es un modelo lineal, ¡a pesar de que hay entradas al cuadrado! Lo que nos interesa es que los pesos no estén al cuadrado.

2. Modelos Lineales. Más ideas.

- Que un modelo sea lineal no está necesariamente ligado al hecho de que la frontera de decisión sea una línea recta.

Ejemplo: en la P1 transformamos nuestras características de entrada, y pudimos resolver satisfactoriamente un problema no linealmente separable con un modelo lineal (regresión lineal).

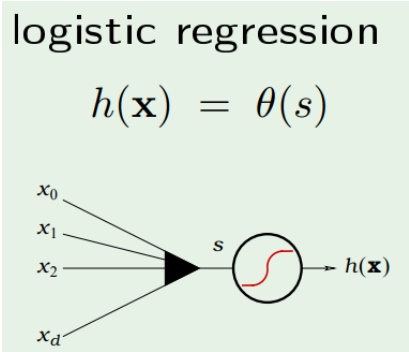
En el fondo, lo único que estábamos haciendo era $y = w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2 + w_4{x_1}^2 + w_5{x_2}^2$ en lugar de $y = w_0 + w_1x_1 + w_2x_2$



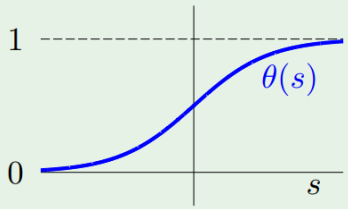
Una frontera de decisión no lineal puede ser producida por un modelo lineal.

2. Modelos Lineales. Más ideas.

- ¿Qué no sería un modelo lineal?
  - Por ejemplo, $y = w_0 \cdot x^{w_1}$
- ¿Cómo es posible que regresión logística sea un modelo lineal si usa una función de activación no lineal?

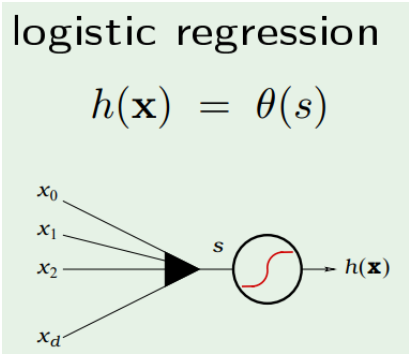


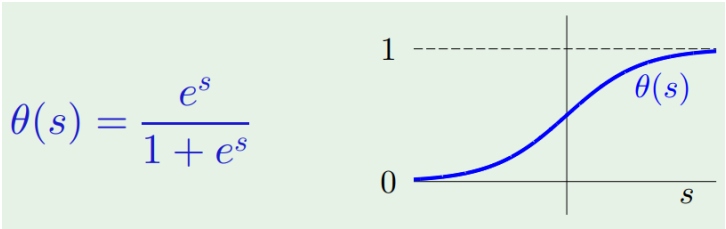
$$\theta(s) = \frac{e^s}{1+e^s}$$



2. Modelos Lineales. Más ideas.

- ¿Cómo es posible que regresión logística sea un modelo lineal si usa una función de activación no lineal?





Porque podemos escribir las predicciones en términos de una función lineal

$$h(x) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} = 0.5 \rightarrow 1 = e^{-\boldsymbol{w}^T\boldsymbol{x}} \rightarrow \ln(1) = \ln(e^{-\boldsymbol{w}^T\boldsymbol{x}}) \rightarrow 0 = -\boldsymbol{w}^T\boldsymbol{x}$$

La salida siempre depende de la suma de los productos de entradas por pesos. No hay ninguna interacción del estilo $w_1x_1 \cdot w_2x_2$, que haría nuestro modelo no lineal.

2. Modelos Lineales. Más ideas.

- ¿Qué pasa si no igualamos la sigmoide a 0.5?

Estaremos asumiendo, a priori, que no hay un 50% de posibilidades de pertenecer a cualquiera de las dos clases. Imaginemos que ese umbral lo establecemos en 0.75: a partir de $h(x) \geq 0.75$ el ejemplo x se asigna a la clase $+1$ y, si es menor, se asigna a la clase -1

$$h(x) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} = 0.75 \rightarrow 1.333 = 1 + e^{-\boldsymbol{w}^T\boldsymbol{x}} \rightarrow 0.333 = e^{-\boldsymbol{w}^T\boldsymbol{x}} \rightarrow$$
$$\ln(0.333) = \ln(e^{-\boldsymbol{w}^T\boldsymbol{x}}) \rightarrow -1.098 = -\boldsymbol{w}^T\boldsymbol{x} \rightarrow 1.098 = \boldsymbol{w}^T\boldsymbol{x}$$

Podemos seguir observando claramente que estamos ante un modelo lineal. Solo cambia el valor en donde se encontrará la frontera de decisión.

2. Modelos Lineales. Más ideas.

- Si regresión logística es lineal y, en cierto sentido, una red neuronal artificial es como la combinación de múltiples unidades logísticas… ¿por qué siempre se dice que una red neuronal es un modelo no lineal?
  - Porque es un modelo no lineal ☺
  - Al encadenar distintas capas de procesado (en donde la salida de una es entrada de otra) sí nos podemos encontrar con expresiones no lineales en los pesos. Ej.: red con dos capas, una única neurona por capa, sin bias y función de activación lineal.

2. Modelos Lineales. Más ideas.

- ¡Ojo!, pero ¿cómo sabemos exactamente si un modelo, sistema o función es lineal?
  - Porque verifica las propiedades de homogeneidad (o escalado) y superposición (o aditividad). En resumen, porque cumplen que $af(x_1) + bf(x_2) = f(ax_1 + bx_2)$
  - Se puede comprobar, por ejemplo que $f(w_1, w_2) = w_1 w_2 x_1$ no es lineal en los pesos, mientras que $f(w_1, w_2) = w_1 + w_2 x_1$ sí lo es.

$$af(w_1, w_2) + bf(w_3, w_4) = f(aw_1 + bw_3, aw_2 + bw_4)$$

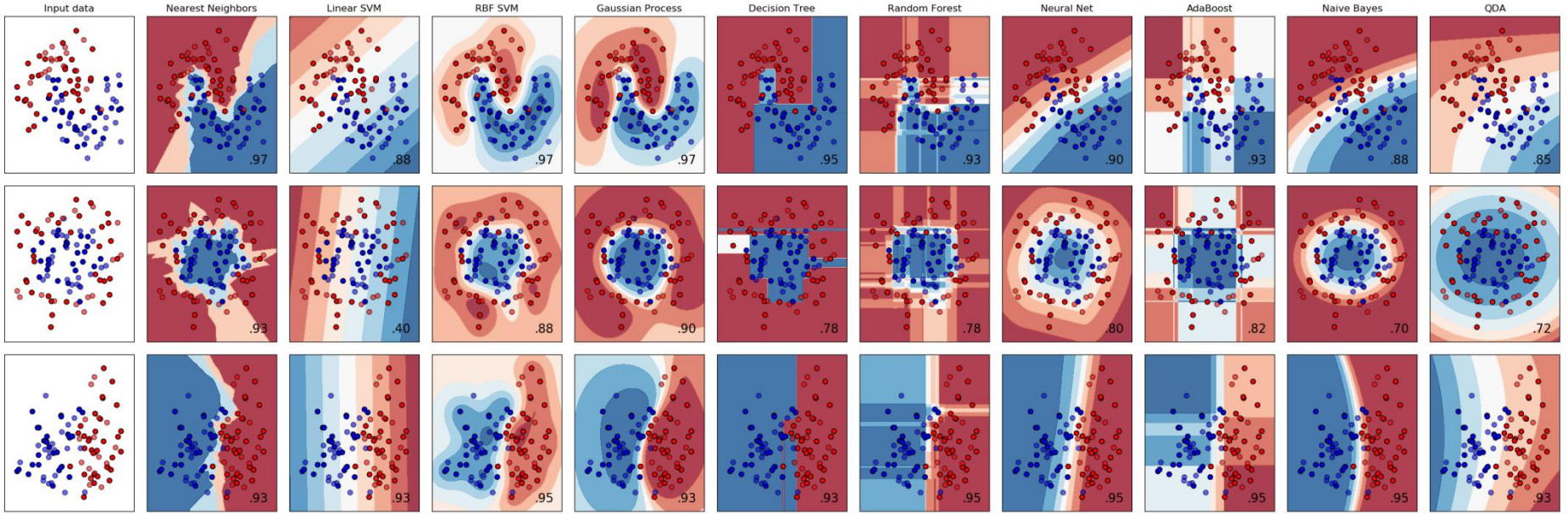
Caso 1) $f(w_1, w_2) = w_1 w_2 x_1$

$$aw_1 w_2 x_1 + bw_3 w_4 x_1 \; != \; (aw_1 + bw_3)(aw_2 + bw_4)x_1$$

Caso 2) $f(w_1, w_2) = w_1 + w_2 x_1$

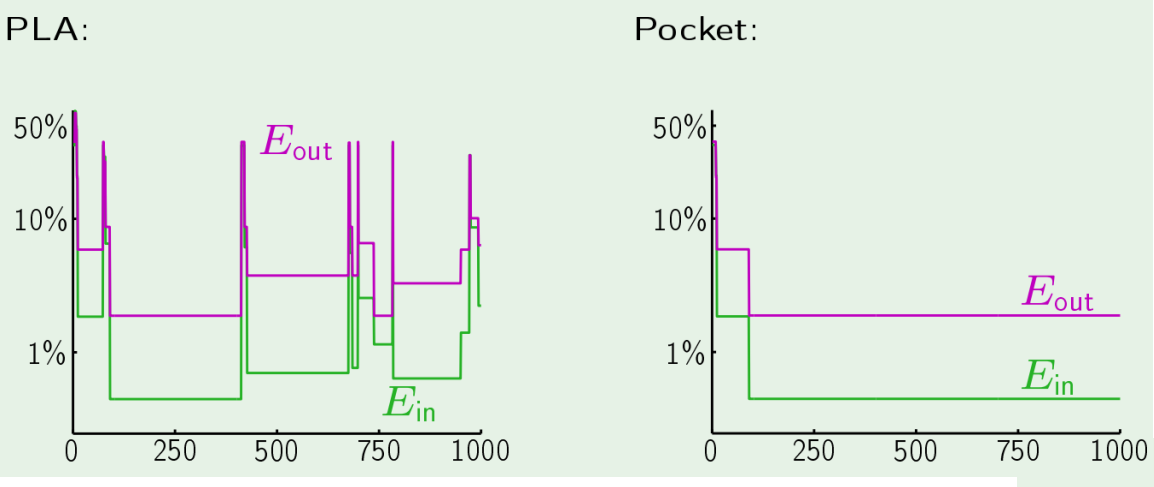
$$a(w_1 + w_2 x_1) + b(w_3 + w_4 x_1) = aw_1 + bw_3 + aw_2 x_1 + bw_4 x_1$$

Interesante referencia para visualizar fronteras de decisión



https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

BONUS

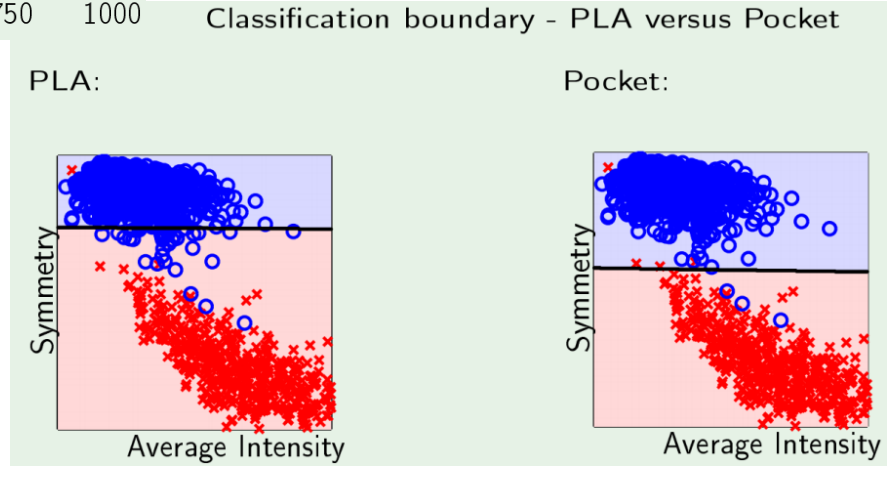


Clasificación de dígitos empleando:

- Regresión lineal
- Regresión logística
- PLA
- PLA-pocket

Y también vais a usar regresión lineal como inicialización para los otros métodos.

http://work.caltech.edu/slides/slides03.pdf slides 7 y 8



BONUS

- The Pocket Algorithm: essentially the pocket algorithm keeps "in its pocket" the best vector solution encountered up the iteration t in PLA

- POCKET ALGORITHM:
  1. Set the pocket weight vector $\hat{w}$ to w(0) of PLA
  2. **for** t=1,…,T **do**
  3. Run PLA for one update to obtain **w**(t+1)
  4. Evaluate E_{in}(**w**(t+1))
  5. If **w**(t+1) is better than $\hat{\mathbf{w}}$ in terms of E_{in}(**w**(t+1)) , set $\hat{\mathbf{w}}$= **w**(t+1)
  6. **Return** $\hat{\mathbf{w}}$

$$\underbrace{\min_{w \in R^{d+1}} \frac{1}{N} \sum_{n=1}^{N} [\![\text{sign}(\mathbf{w}^T \mathbf{x}_n) \neq y_n]\!]}_{E_{in}}$$

- The pocket algorithm has a clear efficiency penalty in point 4.
- But it is guaranteed to get a good solution after a fixed large number of updatings.

BONUS (2.d)

¿Cuál es la dimensión VC del perceptrón?

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} log \frac{4 \cdot ((2N)^{d_{VC}} + 1)}{\delta}}$$

¿Cuál es la cardinalidad de H en test?

$$E_{out}(h) \leq E_{test}(h) + \sqrt{\frac{1}{2N} log \frac{2|H|}{\delta}}$$