EL ALGORITMO PERCEPTRON

④ Probar que PLA converge a un separador lineal en el caso de datos separables. Sea w^* un conjunto optimo de pesos. Sea $w(0)=0$.

a) Llamamos $\rho = \min_{1 \le n \le N} y_n (w^{*T} x_n)$.

Como w^* es un conjunto optimo de pesos, nos proporciona el hiperplano que separa los datos de modo que todos los x_n estarán bien etiquetados y $\rho > 0$.

b) Vemos que $w^T(t) w^* \overset{(1)}{\geq} w^T(t-1) w^* + \rho$.

$w^T(t) w^* = (w^T(t-1) + y(t-1) x(t-1)) w^* =$

usando el algoritmo $= w^T(t-1) w^* + y(t-1) w^{*T} x(t-1) \geq$

por a) $\rightarrow \geq w^T(t-1) w^* + \rho$

Ahora vamos a hacer inducción para verificar que $w^T(t) w^* \overset{(2)}{\geq} t\rho$.

- Caso $t=1$: usando desigualdad (1) y $w(0)=0$

$w^T(1) w^* \geq w^T(1-1) w^* + \rho = w^T(0) w^* + \rho = \rho$

- para t suponemos que es cierto, vemos $t+1$:

$w^T(t+1) w^* \geq w^T(t) w^* + \rho \geq t\rho + \rho = (t+1)\rho$

usando desigualdad (1); hipótesis cierta para t.

c) Mostrar que $\|w(t)\|^2 \overset{(4)}{\leq} \|w(t-\Delta)\|^2 + \|x(t-\Delta)\|^2$.

Antes de nada, sabemos que, como $x(t-\Delta)$ está mal clasificado por $w(t-\Delta)$, se tiene:

$$y(t-\Delta)\cdot(w^T(t-\Delta)\, x(t-\Delta)) \overset{(3)}{\leq} 0$$

Veamos la desigualdad:

$$\|w(t)\|^2 = \|w(t-\Delta) + y(t-\Delta)\, x(t-\Delta)\|^2 =$$

usando el algoritmo →

$$= \|w(t-\Delta)\|^2 + y(t-\Delta)^2 \|x(t-\Delta)\|^2 +$$

$$+ 2y(t-\Delta)(w^T(t-\Delta)\, x(t-\Delta)) \leq$$

por (3) y $y(t-\Delta) \overset{\forall t}{\neq} \Delta$ →

$$\leq \|w(t-\Delta)\|^2 + \|x(t-\Delta)\|^2$$

d) Veamos que $\|w(t)\|^2 \overset{(6)}{\leq} tR^2$ con $R = \max_{1 \leq n \leq N} \|x_n\|$

Caso $t = \Delta$:

$$\|w(\Delta)\|^2 \leq \|w(0)\|^2 + \|x(0)\|^2 = \|x(0)\|^2 \leq R^2$$

por (5) ↓ ; $w(0) = 0$ ↓

Para t suponemos que es cierto, veamos $t+\Delta$:

$$\|w(t+\Delta)\|^2 \leq \|w(t)\|^2 + \|x(t)\|^2 \leq tR^2 + \|x(t)\|^2 \leq$$

por (5) ↓ ; hipótesis ↓ cierta para t

$$\leq tR^2 + R^2 = (t+\Delta)R^2$$

e) Usando b) y d) demostrar que

$$\frac{w^T(t)}{\|w(t)\|}\, w^* \geq \sqrt{t} \cdot \frac{\rho}{R}$$

↑ Sabemos que $w^T(t)\, w^* \geq^{\text{uso (2)}} tp \Rightarrow$

$$\frac{w^T(t)\, w^*}{\|w(t)\|} \geq \frac{tp}{\|w(t)\|} \geq^{\text{uso (6)}} \frac{tp}{\sqrt{t}R} = \frac{\sqrt{t}p}{R}$$

Ahora vemos que $t \leq \frac{R^2 \|w^*\|^2}{p^2}$

↓ Sabemos que $\frac{R^2}{p^2}\left(\frac{w^T(t)\, w^*}{\|w(t)\|}\right)^2 \geq t$

- Asumiendo $\frac{w^T(t)\, w^*}{\|w(t)\|\|w^*(t)\|} \leq^{(7)} 1$, se tiene que

$$\left(\frac{w^T(t)\cdot w^*}{\|w(t)\|}\right)^2 \leq (\|w^*\|)^2 \Rightarrow$$

$$t \leq \frac{R^2}{p^2}\left(\frac{w^T(t)\cdot w^*}{\|w(t)\|}\right)^2 \leq \frac{R^2}{p^2}(\|w^*\|)^2, \text{ es decir,}$$

$$t \leq \frac{R^2 \|w^*\|^2}{p^2}.$$

- (7) es cierto pues $\frac{w^T(t)\, w^*}{\|w(t)\|\|w^*(t)\|}$ es el ángulo θ que forman w y w^*, luego considerando $\cos\theta = \frac{w^T(t)\cdot w^*}{\|w(t)\|\|w^*(t)\|} \leq 1$ se tiene la desigualdad.

Función de crecimiento y punto de ruptura.

(9) Calcula $m_{\mathcal{H}}$ para el modelo de dos círculos concéntricos en $\mathbb{R}^2$.

$\mathcal{H}$ contiene a las funciones que toman valor

$\begin{cases} +1 & \text{en } a^2 \leq x_1^2 + x_2^2 \leq b^2 \\ -1 & \text{en otro caso.} \end{cases}$

Si nos fijamos bien, estamos trabajando con círculos concentros y podemos reducir el problema a una versión equivalente en $\mathbb{R}$.

Para ello, denoto $r = \sqrt{x_1^2 + x_2^2}$, y tendríamos (por hipótesis) que $a \leq r \leq b$.

De este modo hemos llegado al problema visto en clase (example-3). En nuestro caso tenemos:

$$h_{a,b}(r) = \begin{cases} +1 & \text{si } r \in [a,b] \\ -1 & \text{si } r \notin [a,b]. \end{cases}$$

La función de crecimiento es $m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1$.

(Esto es fácil de ver pues tenemos una recta con N valores que separamos por dos "cortes")

Es decir; $m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1$ para nuestro caso de dos círculos concéntricos.

ERROR y RUIDO

① Considerar un modelo que define una hipótesis h, con probabilidad de error μ como aproximación de f (h, f binarias).
Uso h para aproximar una versión ruidosa de f:

$$P(y|x) = \begin{cases} \lambda & y = f(x) \\ 1-\lambda & y \neq f(x). \end{cases}$$

a) Probabilidad de error que comete h al aproximar y.
Sabemos que μ es la probabilidad de cometer un error en los datos que no han sido afectados por el ruido. Claramente, $1-\mu$ será la probabilidad de no cometer error en dichos datos.
Para los datos que sí han sido afectados por el ruido tendremos una probabilidad de $1-\lambda$ de cometer error y probabilidad λ de no cometerlo.
Ya estamos en condiciones de conocer la probabilidad de error que nos piden, pues la probabilidad de error que se comete al aproximar "y" será la probabilidad de no tener error debido al ruido pero sí debido a la aproximación inicial ($f \mu$ por hipótesis) junto con la probabilidad de cometer error debido al ruido pero no debido a la aproximación inicial. Luego dicha probabilidad sería:

$$(1-\lambda)(1-\mu) + \lambda\mu$$

b) Para que valor de λ será h independiente de μ.

Para que h sea independiente de μ, tendremos que establecer un ruido tan grande que los datos se puedan considear aleatorios, careceria de sentido analizar la aproximación a la función f inicial. Viendo la probabilidad de error calculada en el apartado anterior, podemos establecer $\lambda = 0'5$, de modo que quedaría $P_{error} = (1-0'5)(1-\mu) + 0'5\mu =$

$$= 0'5 - 0'5\mu + 0'5\mu = 0'5.$$

Es decir, para $\lambda = 0'5$, tenemos que h será independiente de μ.