Aprendizaje Automático. Cuestiones optativas

Jacinto Carrasco Castillo

9 de marzo de 2016

0.1. El algoritmo Perceptron

Cuestión 1. Probar que PLA finalmente converge a un separador linear en el caso de datos separables. Los siguientes pasos le guiarán a través de la demostración. Sea $w^\star$ un conjunto óptimo de pesos (uno que separa los datos). La idea esencial en esta demostración es mostrar que los pesos $w(t)$ del PLA se alinean cada vez más con $w^\star$ conforme el número de iteraciones avanza. Por simplicidad suponemos $w(0) = 0$

a Sea $\rho = \text{mín}_{1\leq n\leq N}$ y $y_n(w^{\star T} x_n)$. Mostrar que $\rho > 0$

b Mostrar que $w^T(t)w^\star \geq w^T(t-1)w^\star + \rho$ y concluir que $w^T(t)w^\star \geq t\rho$

c Mostrar que $\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$.

d Mostrar por inducción que $\|w(t)\|^2 \leq tR^2$ donde $R = \text{máx}_{1\leq n\leq N}\|x_n\|$

e using (b) y (d), mostrar que

$$\frac{w(t)}{\|w(t)\|}w^\star \geq \sqrt{t}\frac{\rho}{R}$$

y por tanto probar que

$$t \leq \frac{R^2\|w^\star\|^2}{\rho^2}$$

Solución 1. Seguiremos los pasos sugeridos para realizar el ejercicio:

a $w^\star$ separa todos los datos, esto es, $\forall n = 1, \dots, N$, $\texttt{sign(}w^{\star T} x_n\texttt{)} = y_n$. Entonces, tanto si la etiqueta es 1 o -1, se cumple $y_n(w^{\star T} x_n) > 0 \Leftrightarrow \rho = \text{mín}_{1,\dots,N}\, y_n(w^{\star T} x_n) > 0$

b Usaremos inducción para $w^T(t)w^\star \geq tp$:
Para $t = 0, w^T(0)w^\star \geq 0$. Suponemos entonces que para $t-1$, $w^T(t-1)w^\star \geq (t-1)\rho$. Lo probaremos ahora para t:
$w^T(t)w^\star =$ (usando el apartado (a)) $= [w^T(t-1)+y(t)x^T(t)]w^\star \geq w^T(t-1)w^\star + \rho =$ (usando inducción) $= t\rho$

Si escribimos la norma al cuadrado de $w(t)$ como $w^T(t)w(t)$ y usando la definición de $w(t)$ como el ajuste en una iteración, tenemos

$$\|w(t)\|^2 \leq \|w(t-1)\|^2 + 2w^T(t-1)(y(t-1)x(t-1)) + \|x(t-1)\|^2 \leq$$

usando la ayuda sugerida, pues al no estar bien clasificado el dato $x(t-1)$, $y(t-1)w^T(t-1)x(t-1) \leq 0$, y por ser $y(t-1) = \pm 1$ nos queda:

$$\leq \|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$$

$\|w(t)\|^2 \leq tR$, para $t = 0$ es directo. Supuesto para $t-1$,

$\|w(t)\|^2 = [$ por el apartado $(c)] = \|w(t-1)\|^2 + \|x(t-1)\|^2 \leq [$ por ser R el dato con norma máxima $]$

$$\|w(t-1)\|^2 + R^2 \leq [\text{ por inducción }] R^2 + (t-1)R^2 = R^2$$

$\frac{w^T(t)}{\|w(t)\|}w^\star \geq \frac{w^T(t)w^\star}{\sqrt{t}R} = \sqrt{t}\frac{\rho}{R} \Rightarrow$

$$\sqrt{t} = \frac{w^T(t)w^\star R}{\rho\|w(t)\|}$$

Elevando al cuadrado:

$$t \leq \frac{R^2\|w^\star\|}{\rho^2}$$

0.2. Factibilidad del aprendizaje

Cuestión 2. Supongamos que tenemos un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraídos de una función desconocida $f : X \to Y$, donde $X = \mathbb{R}$ e $Y = -1, +1$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = h_1, h_2$ donde h_1 es la función constante igual a $+1$ y h_2 la función constante igual a -1.
Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberadamente la otra hipótesis. Vamos a estudiar cómo estos algoritmos se comportan fuera de la muestra desde un punto de vista determinístico y probabilístico. Suponga en el caso probabilístico que hay una distribución de probabilidad sobre X , y sea $P[f(x) = +1] = p$

a ¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra?

b Asumir desde ahora que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. ¿Es posible que la hipótesis que produce C sea mejor que la hipótesis que produce S?

Solución 2. a No es posible garantizar nada, puesto que la hipótesis que haría S será el valor de la etiqueta que presente un mayor número de individuos en la muestra. Sin embargo, podríamos haber tomado las muestras en una región del espacio de búsqueda donde predominaran puntos de una clase, y sin embargo la distribución general se podría asemejara más a una distribución aleatoria.

b Claro, puesto que, como en el apartado anterior, podríamos haber tomado sólo los 25 puntos con una de las etiquetas.

0.3. Error y Ruido

Cuestión 3. Considerar un modelo (i.e. un recipiente de bolas) que define una hipótesis h cuya probabilidad de error es μ como aproximación de una determinada función determinística f (tanto h como f son binarias). Si usamos la misma h para aproximar una versión ruidosa de f dada por

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

a ¿Cuál es la probabilidad de error que comete h al aproximar y?

b ¿Para qué valor de λ será h independiente de μ?

a Por el enunciado, sabemos que $\mu = E(h) = P[h(x) \neq f(x)]$.