Aprendizaje Automático - Cuestionario 2

Juan Luis Suárez Díaz

27 de abril de 2017

Cuestión 1

Sea X una matriz de números reales de dimensiones $N \times d$, $N > d$, Sea $X = UDV^T$ su descomposición en valores singulares (SVD). Calcular la SVD de X^TX y XX^T en función de la SVD de X. ¿Qué propiedades tienen estas nuevas matrices que no están presentes en X? ¿Qué representa la suma de la diagonal principal de cada una de las matrices producto?

Solución

Supongamos que $X = UDV^T$ es la descomposición SVD de X. Entonces se verifica que U es una matriz ortogonal de $N \times N$ (es decir, $U^T = U^{-1}$) y V es una matriz ortogonal de $d \times d$. Además, D es una matriz diagonal de $N \times d$ de la forma

$$D = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d \\ & O_{N-d\times d} & & \end{pmatrix}$$

Como $X^T = V^{T^T} D^T U^T = VD^TU^T$, y U y V son ortogonales, se tiene:

$$X^TX = VD^TU^TUDV^T = VD^TDV^T = V(D^TD)V^T$$
$$XX^T = UDV^TVD^TU^T = UDD^TU^T = U(DD^T)U^T$$

Donde las matrices D^TD y DD^T son:

$$D^TD = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d \end{pmatrix} O_{d\times N-d} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d \\ & O_{N-d\times d} & & \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_d^2 \end{pmatrix}_{d\times d}$$

$$DD^T = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d \\ & 0_{N-d\times d} & & \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d \end{pmatrix} 0_{d\times N-d} = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \sigma_2^2 & & & & \\ & & \ddots & & & \\ & & & \sigma_d^2 & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{pmatrix}_{N\times N}$$

Como X^TX es de orden $d \times d$, V y V^T son matrices ortogonales de $d \times d$ y D^TD es diagonal de tamaño $d \times d$, $X^TX = V(D^TD)V^T$ es la SVD de la matriz.

Como XX^T es de orden $N \times N$, U y U^T son matrices ortogonales de $N \times N$ y DD^T es diagonal de tamaño $N \times N$, $XX^T = U(DD^T)U^T$ es la SVD de la matriz.

Estas matrices tienen las siguientes propiedades, que en general no son verificadas por X:

- Son cuadradas: X^TX es de orden $d \times d$ y XX^T es de orden $N \times N$.
- Son simétricas, ya que $(X^TX)^T = X^TX^{TT} = X^TX$ (análogo para XX^T).
- Son semidefinidas positivas, ya que las SVDs $X^TX = V(D^TD)V^T$ y $XX^T = U(DD^T)U^T$ son, además una descomposición por congruencia de las matrices (entendiendo como descomposición por congruencia a una de la forma $A = PDP^T$, con D diagonal y P regular), y ambas matrices diagonales tienen todos sus elementos diagonales (valores singulares o ceros) no negativos. Esto es una caracterización de las matrices semidefinidas positivas. Además, si todos los valores singulares son estrictamente positivos, X^TX es además definida positiva, ya que ninguno de sus elementos diagonales es negativo o cero.

Finalmente analizamos la traza del producto. Supongamos que la matriz X viene dada por $X(i,j) = x_{ij}, 1 \leq i \leq N, 1 \leq j \leq d$. Entonces, $X^T(i,j) = x_{ji}, 1 \leq i \leq d, 1 \leq j \leq N$. Aplicando las reglas de multiplicación de matrices, se tiene que:

$$X^TX(i,j) = \sum_{k=1}^{N} X^T(i,k)X(k,j) = \sum_{k=1}^{N} x_{ki}x_{kj}, 1 \leq i \leq d, 1 \leq j \leq d$$

$$XX^T(i,j) = \sum_{k=1}^{d} X(i,k)X^T(k,j) = \sum_{k=1}^{d} x_{ik}x_{jk}, 1 \leq i \leq N, 1 \leq j \leq N$$

En particular, $X^TX(m,m) = \sum_{k=1}^{N} x_{km}^2$ y $XX^T(m,m) = \sum_{k=1}^{d} x_{mk}^2$. Sumando los elementos de las diagonales obtenemos.

$$\sum_{m=1}^{d} X^TX(m,m) = \sum_{m=1}^{d}\sum_{k=1}^{N} x_{km}^2$$

$$\sum_{m=1}^{N} XX^T(m,m) = \sum_{m=1}^{N}\sum_{k=1}^{d} x_{mk}^2$$

Obtenemos por tanto que ambas sumas son iguales y coinciden con $\|X\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{d} x_{ij}^2$, donde $\|\cdot\|_F$ representa la norma matricial de Frobenius, es decir, la norma obtenida del producto escalar usual cuando se ve la matriz como un vector de $\mathbb{R}^{N\times d}$.

En general, siguiendo el mismo razonamiento que el que acabamos de usar se tiene que para cualesquiera matrices A y B multiplicables se verifica que la suma de las diagonales de ambos productos coincide, es decir, $tr(AB) = tr(BA)$. Usando esto, y la descomposición SVD de las matrices producto, obtenemos:

$tr(X^TX) = tr(VD^TDV^T) = tr(D^TDV^TV) = tr(D^TD) = \sum_{i=1}^{d} \sigma_i^2$

Por tanto, obtenemos también que la traza del producto es la suma de los cuadrados de los valores singulares de X. En resumen, hemos obtenido las siguientes igualdades:

$$tr(X^TX) = tr(XX^T) = \sum_{i=1}^{N}\sum_{j=1}^{d} x_{ij}^2 = \|X\|_F^2 = \sum_{i=1}^{d} \sigma_i^2$$

Cuestión 2

Suponga una matriz cuadrada A que admita la descomposición $A = X^TX$ para alguna matriz X de números reales. Establezca una relación entre los valores singulares de la matriz A y los valores singulares de X.

Solución

Si A se puede escribir como el producto de una matriz y su traspuesta, entonces A es necesariamente cuadrada. Supongamos $A \in \mathcal{M}_{n\times n}(\mathbb{R})$. Entonces, $X \in \mathcal{M}_{m\times n}(\mathbb{R})$, para algún $m \in \mathbb{N}$. Supongamos que $X = UDV^T$ es la SVD de X. Entonces, la SVD de A es $A = X^TX = VD^TU^TUDV^T = V(D^TD)V^T$. Si los valores singulares de X son $\sigma_1, \ldots, \sigma_d$, con $d = \text{mín}\{m, n\}$, entonces D es la matriz de dimensión $m \times n$ que tiene como diagonal principal esos valores singulares.

D^TD, por ser D diagonal, es una matriz de dimensiones $n \times n$ y diagonal también. Si $m < n$ los elementos de su diagonal serán $\sigma_1^2, \ldots, \sigma_m^2$ y ceros en el resto de la diagonal. Y si $n \leq m$ todos los elementos de la diagonal serán $\sigma_1^2, \ldots, \sigma_n^2$. En conclusión, lo que tenemos es que los valores singulares de A son los cuadrados de los valores singulares de X, a los que se añaden tantos ceros como sea la diferencia entre n y m (solo cuando $n > m$). Recíprocamente, todo valor singular no nulo de X es la raíz cuadrada de un valor singular no nulo de A. Además, por ser V ortogonal, se tiene que $A = V(D^TD)V^{-1}$, luego D^TD es una diagonalización por semajanza de A y por tanto los valores singulares de A son además valores propios.

Cuestión 3

Definir el problema de optimización con restricciones para encontrar los valores extremos de $f(x, y) = ax + by$ sujeto a la restricción $x^2 + y^2 = r^2$, donde a, b, r son constantes. Definir la Lagrangiana y calcular los valores de x, y y $f(x, y)$ en el óptimo. Discutir las distintas soluciones que se pueden presentar en función de los valores de a, b y r.

Solución

Llamamos $M = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = r^2\}$ al conjunto restricción. En primer lugar, distinguimos los casos degenerados para a y b:

- Si $a = b = 0$, entonces f es la función constantemente 0, y alcanza su máximo y mínimo en cualquier punto de M.
- Si $r = 0$, entonces la $M = \{(0, 0)\}$, luego la restricción se reduce al punto $(0, 0)$, que obviamente maximiza y minimiza f en M, con valor $f(0, 0) = 0$.

Supongamos finalmente $r \neq 0$ y que a y b no son 0 simultáneamente (o equivalentemente, $a^2+b^2 > 0$). La lagrangiana asociada a la función f y a la restricción de M es la aplicación $L : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ dada por:

$$L(x, y, \lambda) = ax + by + \lambda(x^2 + y^2 - r^2)$$

Los extremos de f condicionados por M, por el teorema de Lagrange, proceden de puntos críticos de L. Derivando respecto de cada variable L, obtenemos el siguiente sistema de Lagrange:

$$\begin{cases} a + 2x\lambda = 0 \\ b + 2y\lambda = 0 \\ x^2 + y^2 = r^2 \end{cases}$$

Resolvemos el sistema. Si $\lambda = 0$, entonces obtendríamos que $a = b = 0$, caso que ya hemos descartado previamente, así que podemos suponer $\lambda \neq 0$. Dividiendo por λ, obtenemos de las dos primeras ecuaciones que $x = -a/2\lambda$, $y = -b/2\lambda$. Sustituyendo en la tercera ecuación, tenemos:

$$\frac{a^2}{4\lambda^2} + \frac{b^2}{4\lambda^2} = r^2 \Rightarrow a^2 + b^2 = 4r^2\lambda^2$$

Como estamos suponiendo $r \neq 0$ y $a^2+b^2 > 0$, podemos dividir y tomar raíces cuadradas, obteniendo que $\lambda = \pm\sqrt{\frac{a^2+b^2}{4r^2}} = \frac{\pm\sqrt{a^2+b^2}}{2|r|}$. Es decir, obtenemos las siguientes dos soluciones para el sistema:

$$\begin{cases} \lambda_1 = \frac{+\sqrt{a^2+b^2}}{2|r|} & x_1 = -\frac{a|r|}{\sqrt{a^2+b^2}} \quad y_1 = -\frac{b|r|}{\sqrt{a^2+b^2}} \\ \lambda_2 = \frac{-\sqrt{a^2+b^2}}{2|r|} & x_2 = +\frac{a|r|}{\sqrt{a^2+b^2}} \quad y_2 = +\frac{b|r|}{\sqrt{a^2+b^2}} \end{cases}$$

Por tanto, los posibles valores que minimizan y maximizan f en M son (x_1, y_1) y (x_2, y_2). Evaluamos f en ambos puntos:

$$f(x_1, y_1) = -\frac{a^2|r|}{\sqrt{a^2+b^2}} - \frac{b^2|r|}{\sqrt{a^2+b^2}} = -|r|\sqrt{a^2+b^2} < 0$$

$$f(x_2, y_2) = +\frac{a^2|r|}{\sqrt{a^2+b^2}} + \frac{b^2|r|}{\sqrt{a^2+b^2}} = +|r|\sqrt{a^2+b^2} > 0$$

Por tanto, para cualesquiera $r \neq 0$ y $a, b \in \mathbb{R}$ con $a^2 + b^2 > 0$ se verifica:

- f alcanza su máximo en M en el punto $\left(+\frac{a|r|}{\sqrt{a^2+b^2}}, +\frac{b|r|}{\sqrt{a^2+b^2}}\right)$ y vale $+|r|\sqrt{a^2+b^2}$.
- f alcanza su mínimo en M en el punto $\left(-\frac{a|r|}{\sqrt{a^2+b^2}}, -\frac{b|r|}{\sqrt{a^2+b^2}}\right)$ y vale $-|r|\sqrt{a^2+b^2}$.

Cuestión 4

En regresión lineal con ruido en las etiquetas el error fuera de la muestra para una h dada está dado por:

$$E_{out}(h) = \mathbb{E}_{x,y}[(h(x) - y)^2] = \int\int (h(x) - y)^2 p(x, y) dxdy$$

Mostrar que entre todas las posibles hipótesis, la que minimiza E_{out} está dada por:

$$h^*(x) = \mathbb{E}_y[y|x] = \int y p(y|x) dy$$

Solución

Podemos sumar y restar $\mathbb{E}_y[y|x]$ en la expresión de E_{out}, obteniendo:

$$E_{out}(h) = \int\int (h(x) - \mathbb{E}_y[y|x] + \mathbb{E}_y[y|x] - y)^2 p(x, y) dxdy =$$

$$\int\int \left[(h(x) - \mathbb{E}_y[y|x])^2 + (\mathbb{E}_y[y|x] - y)^2 + 2(h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)\right] p(x, y) dxdy$$

Usando la linealidad de la integral,

$$E_{out}(h) = \int\int (h(x) - \mathbb{E}_y[y|x])^2 p(x, y) dxdy + \int\int (\mathbb{E}_y[y|x] - y)^2 p(x, y) dxdy$$
$$+2\int\int (h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y) p(x, y) dxdy$$

Veamos que el último sumando se anula. Para ello, usamos que $p(x, y) = p(y|x)p(x)$ e integramos respecto de estas probabilidades.

$$\int\int (h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)p(x,y)dxdy = \int\int (h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)p(y|x)dyp(x)dx$$

El término $h(x) - \mathbb{E}_y[y|x]$ es independiente respecto a $p(y|x)dy$, por lo que podemos extraerlo de la primera integral, obteniendo la expresión

$$\int (h(x) - \mathbb{E}_y[y|x]) \int (\mathbb{E}_y[y|x] - y)p(y|x)dyp(x)dx$$

Usando la linealidad en la integral interior, tenemos

$$\int (h(x) - \mathbb{E}_y[y|x]) \left[\mathbb{E}_y[y|x] \int p(y|x)dy - \int yp(y|x)dy\right] p(x)dx$$

Como $p(y|x)$ es una medida de probabilidad, se verifica que $\int p(y|x)dy = 1$. Además, por definición, $\int yp(y|x)dy = \mathbb{E}_y[y|x]$. Sustituyendo en la expresión anterior, obtenemos finalmente

$$\int (h(x) - \mathbb{E}_y[y|x]) \left[\mathbb{E}_y[y|x] - \mathbb{E}_y[y|x]\right] p(x)dx = 0$$

Volviendo a la expresión de E_{out}, por lo que acabamos de ver su expresión se reduce a

$$E_{out}(h) = \int\int (h(x) - \mathbb{E}_y[y|x])^2 p(x,y)dxdy + \int\int (\mathbb{E}_y[y|x] - y)^2 p(x,y)dxdy$$

Tenemos, por tanto, una suma de dos integrales, donde la segunda integral no depende de h, y por tanto minimizar E_{out} se reduce a minimizar la primera integral. Como el integrando de la primera integral es $(h(x) - \mathbb{E}_y[y|x])^2$, que siempre es mayor o igual que 0 y se anula si y solo si $h(x) = \mathbb{E}_y[y|x]$, lo mismo ocurre con toda la integral, por la propiedad de positividad, es decir, $\int\int (h(x) - \mathbb{E}_y[y|x])^2 p(x,y)dxdy = 0 \iff h(x) = \mathbb{E}_y[y|x]$, y por tanto, como el otro término no dependía de h y este alcanza su mínimo en 0, tenemos que h minimiza E_{out} si y solo si $h(x) = \mathbb{E}_y[y|x]$, como queríamos.

Cuestión 5

Escribir la función de Máxima Verosimilitud de una muestra de tamaño N para un problema de clasificación binaria. Además:

a) *Mostrar que la estimación de Máxima Verosimilitud se reduce a la tarea de encontrar la función h que mimimiza*

$$E_{in}(h) = \sum_{n=1}^{N} [y_n = +1] \ln\left(\frac{1}{h(x_n)}\right) + [y_n = -1] \ln\left(\frac{1}{1 - h(x_n)}\right)$$

b) *Para el caso $h(x) = \sigma(w^T x)$ mostrar que minimizar el error de la muestra en el apartado anterior es equivalente a minimizar el error muestral*

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

Solución

a) Podemos ver la función objetivo que queremos aprender, por tratarse de un problema de clasificación binaria, como $f(x) = \mathbb{P}[y = +1|x]$. La probabilidad de que se dé la etiqueta y condicionada a los datos obtenidos será por tanto $P(y|x) = \begin{cases} f(x) & , y = +1 \\ 1 - f(x) & , y = -1 \end{cases}$. Asumiendo que la función hipótesis h aproxima bien a la función objetivo, podemos considerar que la distribución de probabilidad anterior viene dada por $P(y|x) = \begin{cases} h(x) & , y = +1 \\ 1 - h(x) & , y = -1 \end{cases}$. Esta será la verosimilitud para el problema de clasificación con un solo dato, también podemos escribirla como $P(y|x) = h(x)[\![y = +1]\!] + (1 - h(x))[\![y = -1]\!]$. En consecuencia, la verosimilitud para una muestra de N datos, $(x_1, y_1), \ldots, (x_N, y_N)$ será:

$$L(h) = \prod_{n=1}^{N} P(y_n|x_n) = \prod_{n=1}^{N} \left(h(x_n)[\![y_n = +1]\!] + (1 - h(x_n))[\![y_n = -1]\!]\right)$$

Buscamos maximizar $L(h)$. Para simplificar la fórmula, y transformar los productos en sumas, tomamos logaritmos en la igualdad anterior. Como el logaritmo es una función estrictamente creciente, maximizar $L(h)$ equivale a maximizar su logaritmo:

$$\ln(L(h)) = \sum_{n=1}^{N} \ln\left(h(x_n)[\![y_n = +1]\!] + (1 - h(x_n))[\![y_n = -1]\!]\right)$$

Si multiplicamos la expresión anterior por $-1/N$, maximizar la expresión de la verosimilitud será equivalente a minimizar la siguiente expresión:

$$\frac{-1}{N} \ln(L(h)) = \frac{-1}{N} \sum_{n=1}^{N} \ln\left(h(x_n)[\![y_n = +1]\!] + (1 - h(x_n))[\![y_n = -1]\!]\right) =$$

$$\frac{1}{N} \sum_{n=1}^{N} -\ln\left(h(x_n)[\![y_n = +1]\!] + (1 - h(x_n))[\![y_n = -1]\!]\right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \ln\left(\frac{1}{h(x_n)[\![y_n = +1]\!] + (1 - h(x_n))[\![y_n = -1]\!]}\right)$$

Ahora notemos lo siguiente: como cuando $y_n = +1$ el sumando n-ésimo de la expresión anterior vale $\ln\left(\frac{1}{h(x_n)}\right)$, y para $y_n = -1$ cada sumando de la expresión vale $\ln\left(\frac{1}{1-h(x_n)}\right)$, los sumandos de la expresión anterior los podemos reescribir como $\ln\left(\frac{1}{h(x_n)}\right)[\![y_n = +1]\!] + \ln\left(\frac{1}{1-h(x_n)}\right)[\![y_n = -1]\!]$, y por tanto, la expresión anterior es equivalente a:

$$\frac{1}{N} \sum_{n=1}^{N} \left(\ln\left(\frac{1}{h(x_n)}\right)[\![y_n = +1]\!] + \ln\left(\frac{1}{1 - h(x_n)}\right)[\![y_n = -1]\!]\right)$$

Es decir, maximizar la verosimilitud se reduce a minimizar la expresión anterior, la cual es la expresión que íbamos buscando.

b) Si $h(x) = \sigma(w^T x)$, con $\sigma(s) = \frac{e^s}{1+e^s}$, se verifica que $\sigma(-s) = \frac{e^{-s}}{1+e^{-s}} = \frac{1}{1+e^s} = 1 - \frac{e^s}{1+e^s} = 1 - \sigma(s)$ En particular, como consecuencia de esta propiedad, se verifica que $\sigma(yw^T x) = h(x)$, si $y = +1$, y que $\sigma(yw^T x) = 1 - h(x)$, si $y = -1$. Esto nos permite sustituir en la expresión obtenida en el apartado anterior las funciones indicadoras y el propio h por la función σ, obteniendo la expresión:

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{\sigma(y_n w^T x_n)}\right)$$

Finalmente, desarrollando σ concluimos que minimizar la expresión del apartado anterior equivale a minimizar la siguiente:

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

Cuestión 6

Definamos el error en un punto (x_n, y_n) por

$$e_n(w) = \text{máx}(0, -y_n w^T x_n)$$

Argumentar si con esta función de error el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\nu = 1$.

Solución

Como e_n está formado a trozos por dos funciones derivables, se tiene que que e_n es derivable siempre que $y_n w^T x_n \neq 0 \iff w^T x_n \neq 0$ (es decir, siempre que x_n no esté en la frontera del vector de clasificación). Y en tal caso, derivando a trozos, se tiene que el gradiente es:

$$\nabla e_n(w) = \begin{cases} 0 & , \text{ si } y_n w^T x_n > 0 \\ -y_n x_n & , \text{ si } y_n w^T x_n < 0 \end{cases}$$

Por tanto, si la tasa de aprendizaje es 1, la regla de adaptación de pesos del SGD se reduce a:

$$w_{new} = w_{old} - \nabla e_n(w_{old}) = (w_{old} + y_n x_n)[[y_n w^T x_n < 0]]$$

Es decir, la adaptación es $w_{new} = w_{old} + y_n x_n$, siempre que $y_n w^T x_n < 0$, o equivalentemente, siempre que el dato x_n esté mal clasificado. Estamos por tanto, ante la regla de adaptación de pesos del perceptron, luego la función de error nada nos permite interpretar PLA como una versión de gradiente descendente estocástico.

Cuestión 7

Considerar la función de error $E_n(w) = (\text{máx}(0, 1 - y_n w^T x_n))^2$. Argumentar que el algoritmo ADALINE con la regla de adaptación $w_{new} = w + \eta(y_n - w^T x_n) \cdot x_n$ es equivalente a gradiente descendente estocástico (SGD) sobre $\frac{1}{N}\sum_{n=1}^{N} E_n(w)$

Solución

Al ser un algoritmo de gradiente descendente estocástico consideramos la adaptación de pesos puntualmente, luego se trata de movernos en la dirección del gradiente de E_n. E_n es derivable siempre que $1 - y_n w^T x_n \neq 0$ y:

$$\nabla E_n(w) = \begin{cases} 0 & , \text{ si } 1 - y_n w^T x_n < 0 \\ 2(1 - y_n w^T x_n)(-y_n x_n) & , si 1 - y_n w^T x_n > 0 \end{cases}$$

Además, cuando $1 - y_n w^T x_n = 0$ el gradiente tiene límite e igual a 0, luego E_n es derivable para cualquier valor de w. Además, cuando $1 - y_n w^T x_n > 0$ podemos reescribir la expresión del gradiente como $2(1 - y_n w^T x_n)(-y_n x_n) = 2(-y_n x_n + (y_n^2 w^T x_n)x_n) = 2(-y_n + w^T x_n)x_n$ (usando que $y_n^2 = 1$).

Por tanto, la regla de adaptación de pesos del SGD para una tasa de aprendizaje ν se reduce a:

$$w_{new} = w - \nu \nabla E_n(w) = w + 2\nu(y_n w^T x_n)x_n$$

, siempre que $1 - y_n w^T x_n > 0$. Finalmente, si tomamos $\eta = 2\nu$ obtenemos la regla de adaptación de pesos del ADALINE:

$$w_{new} = w + \eta(y_n w^T x_n)x_n$$

Bonus 1

Sea X una matriz $N \times M$, $N > M$, de números reales. ¿Cómo son los valores singulares de las matrices $X, X^T X$ y XX^T y qué relación existe entre ellos?

Solución

Supongamos que $X = UDV^T$ es la descomposición SVD de X, con

$$D = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_M \\ & O_{N-M \times M} & & \end{pmatrix}_{NxM}$$

, con $\sigma_1 \geq \cdots \geq \sigma_M \geq 0$ y U, V ortogonales. Razonando como en los ejercicios 1 y 2 se tiene:

$$X^T X = VD^T U^T UDV^T = V(D^T D)V^T$$
$$XX^T = UDV^T VD^T U^T = U(DD^T)U^T$$

Como V, V^T, U, U^T son matrices ortogonales y $D^T D$ y DD^T son diagonales, las dos descomposiciones son descomposiciones en valores singulares para $X^T X$ y XX^T, respectivamente, y los valores singulares de cada matriz serán los elementos de la diagonal de $D^T D$ y DD^T respectivamente. Calculando ambas matrices como en el ejercicio 1, tenemos que:

$$D^T D = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_M^2 \end{pmatrix}_{M \times M}$$

$$DD^T = \begin{pmatrix} \sigma_1^2 & & & & & & \\ & \sigma_2^2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_M^2 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}_{N \times N}$$

En consecuencia, los M valores singulares de $X^T X$ son los cuadrados de los M valores singulares de X, mientras que los N valores singulares de XX^T son los cuadrados de los M valores singulares de X, a los que se añaden $N - M$ ceros. Por tanto, salvo la diferencia de dimensiones (que se completa con valores singulares nulos), los valores de $X^T X$ y XX^T coinciden y son el cuadrado de los valores singulares de X.

Bonus 2

Sean $\mathcal{H}_1, \ldots, \mathcal{H}_K$, K conjuntos de hipótesis con dimensión de VC finita d_{VC}. Sea $\mathcal{H} = \mathcal{H}_1 \cup \cdots \cup \mathcal{H}_K$ la unión de estos modelos. Mostrar que $d_{VC}(\mathcal{H}) < K(d_{VC} + 1)$

Solución

En primer lugar consideramos los siguientes resultados previos sobre números combinatorios y la dimensión VC:

- Dada una muestra de N puntos y una clase de hipótesis binarias $\mathcal{H}$, $m_{\mathcal{H}}(N)$ representa todas las combinaciones de posibles etiquetas que puede generar $\mathcal{H}$ para N puntos. Claramente, como hay 2^N combinaciones para N etiquetas binarias, siempre se cumple $m_{\mathcal{H}}(N) \leq 2^N$. La dimensión VC de $\mathcal{H}$ es el mayor natural para el que se da la igualdad $m_{\mathcal{H}}(N) = 2^N$.
- Si tenemos las clases de hipótesis $\mathcal{H}_1, \ldots, \mathcal{H}_K$ y $\mathcal{H}$ es la unión de todas, es claro que, a lo sumo, el número de etiquetados diferentes que se pueden generar con $\mathcal{H}$ es como mucho la suma de todos los etiquetados diferentes que se pueden generar con cada hipótesis $\mathcal{H}_i$. Esto muestra que se verifica la siguiente desigualdad: $m_{\bigcup_{i=1}^k \mathcal{H}_\rangle}(N) \leq \sum_{i=1}^k m_{\mathcal{H}_i}(N)$.
- Finalmente, de la teoría sabemos que una cota de $m_{\mathcal{H}}(N)$ es $m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}(\mathcal{H})} \binom{N}{i}$. Sobre números combinatorios, usaremos también la propiedad de simetría, es decir, se verifica $\binom{n}{j} = \binom{n}{n-j}$. Además, se cumple que $\sum_{k=0}^n \binom{n}{k} = 2^n$

Probaremos en primer lugar el siguiente resultado, para la unión de dos clases, y luego generalizaremos por inducción:

Sean $\mathcal{H}_1$ y $\mathcal{H}_2$ dos clases de hipótesis con dimensiones VC d_1 y d_2 respectivamente, y $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$, con dimensión VC d. Entonces se verifica que $d \leq d_1 + d_2 + 1$.

Prueba. Supongamos $N = d_1 + d_2 + 2 > d_1 + d_2 + 1$. Aplicando los resultados previos obtenemos las siguientes desigualdades:

$$m_{\mathcal{H}}(N) \leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i}$$

Usando la propiedad de simetría sobre el segundo sumando, tenemos

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-i}$$

Haciendo el cambio $j = i + N - i$, la expresión queda como sigue

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{j=N-d_2}^{N} \binom{N}{j}$$

Ahora, partiendo de que $N > d_1 + d_2 + 1$, y por tanto $N - d_2 > d_1 + 1$, luego una suma de números combinatorios entre estos índices seá estrictamente positiva, y usando que $\sum_{i=0}^{N} \binom{N}{i} = 2^N$, se tiene:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} < \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=d_1+1}^{N-d_2-1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} = \sum_{i=0}^{N} \binom{N}{i} = 2^N$$

Es decir, para $N > d_1 + d_2 + 1$ hemos obtenido que $m_{\mathcal{H}}(N) < 2^N$, y por tanto necesariamente tiene que verificarse que $d \leq d_1 + d_2 + 1$, pues con más puntos no se pueden generar todos los etiquetados. □

Finalmente probemos el ejercicio por inducción.

- Para $K = 2$, supongamos las clases $\mathcal{H}_1$ y $\mathcal{H}_2$, ambas con dimensión VC d. Por lo que acabamos de probar, la dimensión VC de $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ es $d_{VC}(\mathcal{H}) \leq d + d + 1 = 2d + 1 < 2(d + 1) = K(d + 1)$, probando la desigualdad para la unión de dos clases.
- Supongamos, por inducción, que para $K - 1$ clases $\mathcal{H}_1, \ldots, \mathcal{H}_{K-1}$, todas con dimensión VC d se verifica que $d_{VC}(\bigcup_{i=1}^{K-1} \mathcal{H}_i) < (K - 1)(d + 1)$. Y supongamos que tenemos la clase $\mathcal{H}_K$ con dimensión VC también d. Aplicando el resultado anterior a las dos clases $\bigcup_{i=1}^{K-1} \mathcal{H}_i$ y $\mathcal{H}_k$ (en este caso, $d_1 \leq (K - 1)(d + 1) - 1$ y $d_2 = d$), se tiene:

$$d_{VC}\left(\bigcup_{i=1}^{K} \mathcal{H}_i\right) = d_{VC}\left(\left[\bigcup_{i=1}^{K-1} \mathcal{H}_i\right] \bigcup \mathcal{H}_k\right) \leq ((K - 1)(d + 1) - 1) + d + 1 = (K - 1)(d + 1) + d$$

$$< (K - 1)(d + 1) + d + 1 = K(d + 1)$$

Es decir, $d_{VC}\left(\bigcup_{i=1}^{K} \mathcal{H}_i\right) < K(d + 1)$, como queríamos.