Aprendizaje Automático

Trabajo 2

Cuestiones de Teoría

Nuria Rodríguez Barroso
Universidad de Granada
rbnuria6@gmail.com

Índice

1. Ejercicio 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. ¿Que propiedades tienen estas nuevas matrices que no están presente en X?¿Qué representa la suma de la diagonal principal de cada una de las matrices producto?

La descomposición en valores singulares nos permite descomponer $X = UDV^T$ donde $U \in \mathbb{R}^{N\times N}$ ortogonal, $V \in \mathbb{R}^{d\times d}$ y $D \in \mathbb{R}^{N\times d}$ donde la diagonal se corresponde con los valores singulares de X. Así, D es una matriz de dimensión $N \times d$ compuesta por la matriz diagonal de dimensión $d \times d$ de valores singulares junto con un menor de dimensión $(N - d) \times d$ de ceros debajo.

Análogamente la SVD de X^T será $X^T = (UDV^T) = (V^T)^T D^T U^T = VD^T U^T$, en este caso, $D^T \in \mathbb{R}^{d\times N}$ donde la diagonal se corresponde con los valores singulares de X. Siendo D^T una matriz de dimensión $d \times N$ compuesta por la matriz diagonal de dimensión $d \times d$ de valores singulares junto con un menor de dimensión $d \times (N - d)$ de ceros a la derecha.

Por tanto $X^TX = (VD^TU^T)(UDV^T) = VD^TDV^T$ por ser U ortogonal ($U^TU = Id$) y de la misma forma $XX^T = (UDV^T)(VD^TU^T) = UDD^TU^T$. Así, la descomposición en valores singulares de $X^TX = VD'V^T$ y la descomposición en valores singulares de $XX^T = UD''U^T$.

Luego $X^TX = VD'V^T$ donde D' es una matriz diagonal de dimensión $d \times d$ que contiene en la diagonal los cuadrados de los valores singulares de la matriz X. Por tanto, $X^TX \in \mathbb{R}^{d\times d}$.

Análogamente $XX^T = UD''U^T$ donde D'' es una matriz diagonal de dimensión $N \times N$ que contiene en la diagonal los cuadrados de los valores singulares de la matriz X en los primeros d elementos y en el resto son 0. Por tanto, $XX^T, \in \mathbb{R}^{N\times N}$.

Las propiedades que presentan las matrices XX^T y X^TX que no están presentes en X es que son matrices cuadradas, simétricas y diagonalizables. La primera propiedad la hemos visto directamente con las dimensiones resultantes tras la multiplicación de las matrices ($A \in \mathbb{R}^{a\times b}, B \in \mathbb{R}^{b\times c} \rightarrow AB \in \mathbb{R}^{a\times c}$). La justificación de la diagonalización es trivial pues hemos obtenido una diagonalización directamente. La justificación de la simetría se reduce a la propiedad de las matrices que asegura que A^TA es una matriz simétrica ($e_{ij} = e_{ji}$) $\forall i, j \in \{1, ..., dim(A^TA)\}$. Esta propiedad se deduce fácilmente del producto de matrices pues si observamos el producto de una matriz y su transpuesta:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{M1} & \dots & \dots & a_{MN} \end{pmatrix} \times \begin{pmatrix} a_{11} & a_{21} & \dots & a_{M1} \\ a_{12} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{1N} & \dots & \dots & a_{MN} \end{pmatrix}$$

Obtenemos que en la matriz producto, el elemento

$$a_{ij} = \sum_{n=1}^{N} a_{in}a'_{nj} = \sum_{n=1}^{N} a'_{ni}a_{jn} = \sum_{n=1}^{N} a_{jn}a'_{ni} = a_{ji}$$

donde a, a' representan los elementos de A, A^T respectivamente. En la segunda igualdad hemos usado que las dos matrices que estamos multiplicando son transpuestas, esto es $a_{ij} = a'_{ij}$.

En cuanto a la suma de la diagonal principal (traza) de cada una de las matrices producto, podemos asegurar que es la misma que la traza de las matrices D' y D''. Esto se justifica con un sencillo argumento de diagonalización que asegura que la diagonalización por congruencia no varía la traza de la función. Luego se corresponde con la suma de los cuadrados de valores singulares de X.

2. Ejercicio 2:

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

Supongamos $A \in \mathbb{R}^{a\times a}$ entonces $X^T \in \mathbb{R}^{a\times b}$ y $X \in \mathbb{R}^{b\times a}$. Entonces, si consideramos la descomposición en valores singulares $X = UDV^T$, tenemos que $A = (UDV^T)^T(UDV^T) = VD^TU^TUDV^T = VD^TDV^T$, luego hemos obtenido una descomposición de A como $A = VD'V^T$ donde $D' = D^TD$.

Luego, en vista de la descomposición que hemos obtenido distinguimos dos casuísticas:

- $a > b$: Tenemos que nos encontramos con el caso del ejercicio anterior, obtenemos que en la descomposición que ya hemos desarrollado para $A = VD'V^T$ con $D' = D^TD$ donde D' es una matriz diagonal de dimensión $a \times a$. Luego, los valores de la diagonal de D' son los mismo valores de la diagonal del menor principal de dimensión $a \times a$ de D, que es el mismo que el de D^T. Así, con el mismo argumento de diagonalización por congruencia que hemos utilizado anteriormente obtenemos que los valores singulares de A se corresponden con los valores singulares de X.

  Podemos ver un ejemplo muy sencillo con matrices directamente diagonales en el que se comprueba lo anteriormente dicho.

$$\begin{pmatrix}1&0&0\\0&4&0\\0&0&1\end{pmatrix} = \begin{pmatrix}1&0&0&0&0\\0&2&0&0&0\\0&0&1&0&0\end{pmatrix} \times \begin{pmatrix}1&0&0\\0&2&0\\0&0&1\\0&0&0\\0&0&0\end{pmatrix}$$

- $a \leq b$: En este caso obtenemos, que según la descomposición en valores singulares de A anteriormente desarrollada la matriz D' se corresponde con una matriz de orden $b \times b$, por tanto, los valores de la diagonal de D' se corresponden con los valores de las diagonales de los menores principales de dimensión $a \times a$ de D y D^T y $(b - a)$ ceros. Por tanto, con el mismo argumento de congruencia, los valore singulares de A son los valores singulares de D' que son los valores singulares de X, junto con $(b - a)$ ceros.

  Podemos ver un ejemplo muy sencillo con matrices directamente diagonales en el que se comprueba lo anteriormente dicho.

$$\begin{pmatrix}1&0&0&0&0\\0&4&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix} = \begin{pmatrix}1&0&0\\0&2&0\\0&0&1\\0&0&0\\0&0&0\end{pmatrix} \times \begin{pmatrix}1&0&0&0&0\\0&2&0&0&0\\0&0&1&0&0\end{pmatrix}$$

3. Ejercicio 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.

Sean $k, m, n \in \mathbb{N}$ verificando $k + m = n$. Sea $G \subset \mathbb{R}^n$, $g : G \to \mathbb{R}^m$, $g \in \mathcal{C}^1$ tal que el rango $(J_g(x)) = m \quad \forall x \in G$ y $f : G \to \mathbb{R}$ derivable. Sea $M = \{x \in G : g(x) = 0\}$ la variedad de dimensión k determinada por g.

En este contexto, definimos la Langrangiana $L : G \times \mathbb{R}^m \to \mathbb{R}$ como:

$$L(x, \lambda) = f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + ... + \lambda_m g_m(x)$$

Ahora, haciendo uso del *Teorema de Lagrange* sabemos que los puntos críticos de f condicionados a M se corresponden con los puntos críticos de la función Lagrangiana L.

Aplicamos esta teoría para resolver nuestro ejercicio:

1. Definimos $f(x,y) = ax + by$.
2. Definimos $g(x,y) = x^2 + y^2 - r^2$
3. Definimos $L(x,y,\lambda) = ax + by + \lambda(x^2 + y^2 - r^2)$
4. Hallamos los puntos críticos de la función L.
5. El gradiente de L viene dado por $\nabla L(x,y,\lambda) = (a + 2x\lambda, b + 2y\lambda, x^2 + y^2 - r^2)$.
6. Resolvemos el sistema $\nabla L(x,y,\lambda) = (0,0,0)$

$$\begin{cases} a + 2x\lambda = 0 \\ b + 2y\lambda = 0 \\ x^2 + y^2 - r^2 = 0 \end{cases} \iff \begin{cases} a + 2x\lambda = 0 \\ b + 2y\lambda = 0 \\ x^2 + y^2 = r^2 \end{cases} \iff \begin{cases} x = \frac{-a}{2\lambda} \\ y = \frac{-b}{2\lambda} \\ x^2 + y^2 = r^2 \end{cases} \iff (\frac{-a}{2\lambda})^2 + (\frac{-b}{2\lambda})^2 = r^2 \iff$$

$$\iff \frac{a^2 + b^2}{4\lambda^2} = r^2 \iff a^2 + b^2 = 4\lambda^2 r^2 \iff \lambda^2 = \frac{a^2 + b^2}{4r^2} \iff \lambda = \frac{\pm\sqrt{a^2 + b^2}}{2|r|}$$

7. Hallamos (x,y) puntos críticos despejando del sistema de ecuaciones anterior obteniendo dos soluciones, una para cada valor de λ:

$$\begin{cases} a + 2x\frac{\pm\sqrt{a^2+b^2}}{2|r|} = 0 \\ b + 2y\frac{\pm\sqrt{a^2+b^2}}{2|r|} = 0 \\ x^2 + y^2 - r^2 = 0 \end{cases} \iff \begin{cases} 2x\frac{\pm\sqrt{a^2+b^2}}{2|r|} = -a \\ 2y\frac{\pm\sqrt{a^2+b^2}}{2|r|} = -b \end{cases} \iff \begin{cases} x_1 = \frac{-a|r|}{\sqrt{a^2+b^2}}, & y_1 = \frac{-b|r|}{\sqrt{a^2+b^2}} \\ x_2 = \frac{a|r|}{\sqrt{a^2+b^2}}, & y_2 = \frac{b|r|}{\sqrt{a^2+b^2}} \end{cases}$$

8. Evaluamos en la función f para determinar cuál de ellos es el máximo y cuál es el mínimo:

$$\begin{cases} f(x_1, y_1) = \frac{-|r|(a^2+b^2)}{\sqrt{a^2+b^2}} \leq 0 \\ f(x_2, y_2) = \frac{|r|(a^2+b^2)}{\sqrt{a^2+b^2}} \geq 0 \end{cases}$$

Luego el punto en el que se alcanza el máximo es (x_2, y_2) y el valor máximo de la función f es

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

y el punto en el que se alcanza el mínimo es (x_1, y_1) y el valor mínimo de la función f es

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

Para todos los cálculos anteriores hemos considerado, $a, b, r \neq 0$. Veamos qué ocurriría en estos casos:

- Si $r = 0$, el único punto sujeto a la restricción sería el $(x_0, y_0) = 0$, luego él mismo sería el único punto crítico.
- Si $a = b = 0$ entonces obtendríamos que nuestra función sería la función constantemente cero, por lo tanto todos los puntos que verifican la restricción son puntos críticos.
- Si $a, b, r < 0$ no cambiaría nada pues la r está elevada al cuadrado en la restricción y las soluciones se expresan en función de a y b al cuadrado.
- En el resto de los casos se pueden realizar los cálculos realizados y llegar a las soluciones obtenidas.

4. Ejercicio 4:

En la regresión lineal con ruido en las etiquetas, el error fuera de la muestra para una h dada esta 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:

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

La idea para probar este ejercicio es sumar y restar dentro del cuadrado de la integral $\mathbb{E}_y[y|x]$, lo cual al estar sumando y restando no afecta al resultado pero nos va a facilitar la minimización de la función. Luego la función que vamos a minimizar quedaría:

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

Tomando $a = (h(x) - \mathbb{E}_y[y|x])$ y $b = (\mathbb{E}_y[y|x] - y)^2$ podemos desarrollar la igualdad notable de la forma $(a + b)^2 = a^2 + b^2 + 2ab$, por lo que la integral nos quedaría:

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

Luego, hemos conseguido convertir la función que teníamos dentro de la integral en un sumando de tres funciones, por tanto, podemos dividir la integral en la suma de tres integrales de la forma:

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

Con el objetivo de simplificar la ecuación para minimizarla, veamos que el tercer sumando vale siempre 0. Para ello, utilizaremos un resultado básico de probabilidad que se relaciona la probabilidad compuesta de x e y con la probabilidad condicionada de y a x de la siguiente forma: $p(x, y) = p(y|x)p(x)$. Haciendo uso de esto el tercer sumando nos queda de la forma:

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

Según el *Teorema de Fubini* sabemos que podemos intercambiar el orden de las integrales obteniendo que:

$$2\int\int (h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y) p(y|x) p(x) dxdy = 2\int\int (h(x) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y) p(y|x) p(x) dydx$$

Así, seguimos desarrollando nuestro razonamiento por el segundo lado de la igualdad obteniendo:

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

Desarrollamos ahora la integral respecto de y obteniendo:

$$\int (\mathbb{E}_y[y|x] - y)p(y|x)dy = \mathbb{E}_y[y|x] \int p(y|x)dy - \int yp(y|x)dy = \mathbb{E}_y[y|x] - \mathbb{E}_y[y|x] = 0$$

Para la resolución de las dos integrales hemos usado que p es una probabilidad, luego integra 1 por estar bien definida y por definición $\mathbb{E}_y[y|x] = \int yp(y|x)dy$.

Ahora, sustituyendo en el paso anterior obtenemos:

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

Luego, hemos obtenido que nuestra ecuación inicial la podemos ver como:

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

Por tanto, el problema de minimización se reduce a minimizar esta ecuación, para ello minimizamos el primer sumando (dado que el segundo no depende de h), centrémonos en el segundo sumando y obtenemos:

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

$$= \int (h(x) - \mathbb{E}_y[y|x])^2 p(x)dx$$

que claramente se minimiza cuando

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

por lo que hemos obtenido lo que buscábamos.

5. Ejercicio 5:

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

Definimos la función de Máxima Verosimilitud como una expresión de la probabilidad conjunta de probabilidades de una muestra de tamaño N, $\mathcal{D} = (x_1, y_n), ..., (x_N, y_N)$ de instancias independientes e idénticamente distribuidas *(iid)* como:

$$\prod_{n=1}^{N} p(y_n|x_n)$$

donde $p(y_n|x_n)$ representa la probabilidad en un punto. La probabilidad en un punto en un problema de clasificación binaria se define de la siguiente manera:

$$p(y|x) = \begin{cases} f(x), & y = +1 \\ 1 - f(x), & y = -1 \end{cases}$$

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

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

La estimación de la Máxima Verosimilitud se reduce al problema de maximizar la función de Máxima Verosimilitud cuando tomamos $h = f$, si hacemos esto obtenemos que debemos maximizar:

$$\prod_{n=1}^{N} p(y_n|x_n)$$

donde

$$p(y|x) = \begin{cases} h(x), & y = +1 \\ 1 - h(x), & y = -1 \end{cases}$$

En probabilidad es muy común el uso de la propiedad del logaritmo para la obtención de puntos críticos de las funciones. Dado que el logaritmo es monótonamente creciente, el encontrar un máximo de la función $g(x)$ es equivalente a encontrar un máximo de la función $ln(g(x))$, así nuestro problema es equivalente a maximizar

$$ln\left(\prod_{n=1}^{N} p(y_n|x_n)\right)$$

Haciendo uso de las funciones indicadoras podemos reescribir el problema de maximización de manera equivalente:

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

Claramente, el problema de maximizar una función creciente es equivalente al problema de minimizar su opuesto. Así, el problema que contemplábamos al principio se reduce a minimizar:

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

Aplicando una propiedad de los logaritmos que asegura que $a*ln(b)=ln(b^a)$ obtenemos que:

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

Ahora bien, veamos por distinción de casos que:

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

Distinguiendo los casos obtenemos que:

1. Si $y_n=1$ entonces:

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

2. Si $y_n=-1$ entonces:

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

Luego se verifica la igualdad anterior. Ahora, usando la propiedad de los logaritmos que nos asegura que $log(a\times b)=log(a)+log(b)$ obtenemos:

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

Por lo que hemos obtenido lo que deseábamos demostrar.

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

$$E_{in}(w)=\frac{1}{N}\sum_{n=1}^{N}\ln(1+e^{-y_nw^Tx_n})$$

Recordamos que la definición de $\sigma(w^T x) = \frac{e^{w^T x}}{1+e^{w^T x}} = \frac{1}{1+e^{-w^T x}}$. Así, es fácilmente comprobable que la función logística verifica $\sigma(-s) = 1 - \sigma(s)$. Veámoslo:

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

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

Ahora bien, si tomamos $h(x) = \sigma(w^T x)$, obtenemos que la función de probabilidad en un punto se puede resumir en $p(y|x) = \sigma(yw^T x)$. Esto es fácilmente comprobable pues:

$$p(y|x) = \sigma(yw^T x) = \begin{cases} \sigma(w^T x) = h(x), & y = +1 \\ \sigma(-w^T x) = 1 - h(x), & y = -1 \end{cases}$$

Por tanto, podemos reescribir la función de verosimilitud de la forma:

$$\prod_{n=1}^{N} p(y_n|x_n) = \prod_{n=1}^{N} \sigma(y_n w^T x_n)$$

Claramente, minimizar el error producido al dar un valor a la función h se traduce en un problema de maximización de la verosimilitud asociada a esta nueva función h. Aplicando el razonamiento del apartado anterior, obtenemos que el problema se reduce a calcular el máximo de:

$$\sum_{n=1}^{N} ln(\sigma(y_n w^T x_n))$$

Dado que el cálculo de un máximo en una función creciente no varía por la multiplicación por constantes (es proporcional, sigue siendo una función creciente) tenemos que nuestro problema de maximización es análogo a:

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

Con el mismo razonamiento del apartado a) transformamos el problema en un problema de minimización de:

$$-\frac{1}{N} \sum_{n=1}^{N} ln(\sigma(y_n w^T x_n))$$

Y por la misma propiedad de los logaritmos obtenemos

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

6. Ejercicio 6:

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

$$e_n(w) = max(0, -y_n W^T x_n)$$

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

Recordamos que la regla de adaptación del algoritmo PLA se define como:

$$w_{new} = w_{old} + x_n y_n$$

Análogamente, en el SGD, la regla de adaptación que se lleva a cabo es:

$$w_{new} = w_{old} - \nu \nabla e_n(w_{old})$$

El algoritmo PLA utiliza la regla de adaptación en un único paso, esto es, cuando encuentra un punto mal clasificado, calcula el nuevo vector de características según la regla de adaptación y comienza a recorrer otra vez los puntos (en otro orden si lo hacemos aleatorio) hasta encontrar otro punto mal clasificado y así hasta que no haya más puntos clasificados o se verifique algún criterio de parada impuesto (por ejemplo: número máximo de iteraciones). Por su parte, el algoritmo SGD recorre el vector de puntos (en orden aleatorio), y para cada punto realiza la regla de adaptación anteriormente descrita.

Veamos que si sustituimos ∇e_n con la definición del error que nos proporcionan y la tasa de aprendizaje por $\nu = 1$ obtenemos que la regla de adaptación del algoritmo SGD puede interpretarse como el algoritmo PLA.

Para empezar, para facilitar los cálculos del gradiente consideramos la función del error (expresada como una función máximo) como una función definida a trozos en función de los valores que tome la función inicial. Así obtenemos

$$e_n(w) = \begin{cases} 0, & y_n w^T x_n \geq 0 \\ -y_n w^T x_n, & y_n W^T x_n < 0 \end{cases}$$

Claramente e_n es una función continua luego podemos calcular su gradiente. Calculando el gradiente de e_n obtenemos

$$e_n(w) = \begin{cases} 0, & y_n w^T x_n \geq 0 \\ -y_n x_n, & y_n W^T x_n < 0 \end{cases}$$

Por tanto, si sustituimos en la regla de adaptación del SGD obtenemos:

$$w_{new} = w_{old} - \begin{cases} 0, & y_n w^T x_n \geq 0 \\ -y_n x_n, & y_n w^T x_n < 0 \end{cases}$$

Para que esta regla se corresponda con la regla seguida por el algoritmo del perceptrón, tendríamos que comprobar que que se verifique $y_n w^T x_n \geq 0$ se corresponde con que (x_n, y_n) esté bien clasificado. Así, obtendríamos que

$$w_{new} = \begin{cases} \mathrm{w}_{old}, \text{ si el punto está bien clasificado} \\ \mathrm{w}_{old} + x_n y_n, \text{ si el punto está mal clasificado} \end{cases}$$

Lo cual se corresponde exactamente con la regla del perceptrón.

Para probar la propiedad anterior, simplemente notar que si $y_n w^T x_n \geq 0$ entonces tenemos que $sign(y_n) = sign(w^T x_n)$, que es la definición de que el punto esté bien clasificado en el perceptrón. Pues si $sign(y_n) \neq sign(w^T x_n)$ entonces tendríamos claramente $y_n w^T x_n < 0$.

Por lo tanto, hemos obtenido que la regla obtenida si utilizamos en el SGD la expresión del error dada con la tasa de aprendizaje $\eta = 1$ es igual a la regla del perceptrón.

Nota: He incluido el caso de $y_n w^T x_n = 0$ como caso de punto bien clasificado pues un punto por el que justo pase la recta se

7. Ejercicio 7:

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

La principal diferencia entre el algoritmo ADALINE y el PLA, y la cual constituye la mejora sobre el PLA es que en una etapa del algoritmo, este aplica la regla de adaptación teniendo en cuenta a todo punto mal clasificado según el criterio del ADALINE, mientras que el PLA solo modifica el vector de pesos con el punto mal clasificado encontrado. El criterio no es el mismo que el del algoritmo PLA pues el ADALINE es un método basado en distancias y por tanto el criterio de mal clasificación se define como: "un punto está mal clasificado si $y_n w^T x_n < 1$".

Por tanto, teniendo en cuenta estas consideraciones, el problema a tratar es muy similar al ejercicio anterior. Entonces, análogamente redefinimos la función de la forma:

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

Por tanto, el gradiente sería:

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

Así, sustituyendo en la regla de adaptación del gradiente descendente estocástico obtenemos sobre $\frac{1}{N}\sum_{n=1}^{N} E_n(w)$, obtenemos que:

$$w_{new} = w_{old} - \nu \nabla \frac{1}{N}\sum_{n=1}^{N} E_n(w_{old}) = w_{old} - \nu \frac{1}{N}\sum_{n=1}^{N} \nabla E_n(w_{old}) = w_{old} - \nu \begin{cases} 0, & y_n w^T x_n < 1 \\ -2(1 - y_n w^T x_n)(y_n x_n), & y_n w^T x_n \geq 1 \end{cases}$$

Por lo tanto, como el ADALINE solo aplica la regla de adaptación a los puntos mal clasificados (aquellos en los que $y_n w^T x_n < 1$), se correspondería con aplicar

$$w_{new} = w_{old} - \frac{2\nu}{N}\nu(1 - y_n w^T x_n)(-y_n x_n) = w_{old} + \frac{2\nu}{N}\nu(y_n - y_n^2 w^T x_n)x_n$$

Notando que como $y_n = \pm 1$, entonces $y_n^2 = 1$ y tomando $\eta = \frac{2\nu}{N} < 1$, obtenemos que la regla de adaptación del algoritmo SGD se correspondería con

$$w_{new} = \begin{cases} \mathrm{w}_{old} + \eta(y_n - w^T x_n)x_n \text{ cuando } y_n w^T x_n \geq 1 \\ \mathrm{w}_{old} \text{ cuando } y_n w^T x_n < 1 \end{cases}$$

Lo cual se corresponde exactamente con la regla de adaptación del algoritmo ADALINE.

8. BONUS

8.1. Ejercicio 1:

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

Dado que X es una matriz de orden $N \times M$ con $N > M$, rápidamente obtenemos las dimensiones de las matrices X^TX y XX^T. Claramente XX^T tiene dimensión $N \times N$ mientras que X^TX tiene dimensión $M \times M$.

Haciendo uso de la descomposición en valores singulares de $X = UDV^T$, obtenemos que los valores singulares de X son aquellos que conforman la matriz "diagonal" D. Dado que D es de dimensión $N \times M$, luego tendrá M valores singulares.

Repitiendo el proceso del ejercicio 1, obtenemos que las descomposiciones en valores singulares de las otras dos matrices son:

- $X^TX = UD''V^T$ donde D'' es una matriz de dimensión $N \times N$, la cual contiene los M valores singulares de X al cuadrado y $(N - M)$ ceros en la diagonal. Por tanto, los valores singulares de X^TX, haciendo uso nuevamente de la caracterización de la diagonalización por congruencia son los M valores singulares de X al cuadrado, junto con más ceros.
- $XX^T = UD'V^T$ donde D' es una matriz de dimensión $M \times M$, la cual contiene todos los valores singulares de X al cuadrado. Por tanto, por el mismo razonamiento del punto anterior, los valores singulares de XX^T son los valores singulares de X al cuadrado.

8.2. Ejercicio 3:

Suponga que tiene un conjunto de datos con 100 puntos. Suponga que tiene 100 modelos cada uno con dimensión de VC igual a 10. Dispone de 25 puntos adicionales para validación. Elige el modelo que produce un menor error de validación, que resulta ser de 0.25.

a) Dar una cota para E_{out} del modelo seleccionado.

Para dar una cota del error E_{out}, en función del error de validación, esto es, del error obtenido en unos datos test podemos utilizar la desigualdad de Hoeffdings. Esta desigualdad nos asegura que, siendo el error de validación E_{test} tenemos:

$$P(|E_{test} - E_{out}| > \epsilon) \geq 2e^{-2N\epsilon^2}$$

Así, fijando un $\lambda \in (0, 1)$, con una confianza del $1 - \lambda$ podemos asegurar que $E_{out} \leq E_{test} + \epsilon$. Tomando $\lambda = 2e^{-2N\epsilon^2}$ y resolviendo obtenemos que:

$$\lambda = 2e^{-2N\epsilon^2} \rightarrow \frac{\lambda}{2} = e^{-2N\epsilon^2} \rightarrow -2N\epsilon^2 = \ln\left(\frac{\lambda}{2}\right) \rightarrow \epsilon = \sqrt{\frac{1}{2N} ln\left(\frac{2}{\lambda}\right)}$$

Así, tomando $E_{test} = 0{,}25$, $N = 25$ (puntos de test) obtenemos:

$$E_{out} \leq 0{,}25 + \sqrt{\frac{1}{50} ln \left(\frac{2}{\lambda} \right)}$$

una cota en función de λ.

b) Suponga que ha entrenado el modelo sobre todos los datos y selecciona el modelo con menor valor de E_{in} que resulta ser de 0.15. Dar una cota para E_{out} en este caso.

En este caso, utilizamos la desigualdad que afirma que:

$$E_{out} \leq E_{in} + \sqrt{\frac{8}{N} \ln \left(\frac{4((2N)^{VC} + 1)}{\lambda} \right)}$$

Tomando los valores $E_{in} = 0{,}15$, $VC = 10$ y $N = 100$, que son los puntos del conjunto (no contamos los 25 puntos adicionales para validación, pues damos por hecho que el E_{in} ha sido calculado solo con los otros 100 puntos, si no tomamos $M = 125$) obtenemos:

$$E_{out} \leq E_{in} + \sqrt{\frac{8}{100} \ln \left(\frac{4((200)^1 0 + 1)}{\lambda} \right)}$$

una cota en función de λ.