Aprendizaje Automático: Cuestionario 2

Anabel Gómez Ríos

15 de mayo de 2016

1. Cuestiones

Pregunta 1. Sean **x** e **y** dos vectores de observaciones de tamaño N. Sea

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

la covarianza de dichos vectores, donde $\overline{z}$ representa el valor medio de los elementos de **z**. Considere ahora una matriz X cuyas columnas representan vectores de observaciones. La matriz de covarianzas asociada a la matriz X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Defina la expresión matricial que expresa la matriz $cov(X)$ en función de la matriz X.

Vamos a llamar $X = (x_1, x_2, ..., x_M)$ con x_i, $i = 1...M$ vectores columna. Entonces

$$cov(X) = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & ... & cov(x_1, x_M) \\ ... & ... & ... & ... \\ cov(x_M, x_1) & cov(x_M, x_2) & ... & cov(x_M, x_M) \end{pmatrix}$$

Ahora, vamos a desarrollar la igualdad dada para utilizarla en esta matriz:

$$cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N}(x_i y_i - x_i\bar{y} + \bar{x}y_i + \bar{x}\bar{y}) =$$

Ahora, $\bar{x}$ y $\bar{y}$ son independientes de i y los podemos sacar fuera de la suma y separar la suma, luego

$$= \bar{x}\bar{y} - \bar{x}\frac{1}{N}\sum_{i=1}^{N} y_i - \bar{y}\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} x_i y_i =$$

y como $\bar{x} = \sum_{i=1}^{N} x_i$ e igualmente con $\bar{y}$, y $\sum_{i=1}^{N} x_i y_i = x^T y$, podemos escribir:

$$\bar{x}\bar{y} - \bar{x}\bar{y} - \bar{y}\bar{x} + \frac{1}{N}x^T y = -\bar{x}\bar{y} + \frac{1}{N}x^T y$$

con lo que hemos llegado a que $cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = -\bar{x}\bar{y} + \frac{1}{N}x^T y$ y podemos utilizar esto en cada elemento de la matriz anterior:

$$\begin{pmatrix} -\bar{x}_1^2 + \frac{1}{N}x_1^T x_1 & -\bar{x}_1\bar{x}_2 + \frac{1}{N}x_1^T x_2 & ... & -\bar{x}_1\bar{x}_M + \frac{1}{N}x_1^T x_M \\ ... & ... & ... & ... \\ -\bar{x}_M\bar{x}_1 + \frac{1}{N}x_M^T x_1 & -\bar{x}_M\bar{x}_2 + \frac{1}{N}x_M^T x_2 & ... & -\bar{x}_M^2 + \frac{1}{N}x_M^T x_M \end{pmatrix}$$

Como en cada elemento tenemos dos sumandos bien diferenciados, los vamos a separar en dos matrices distintas, sumando:

$$\begin{pmatrix} -\bar{x}_1^2 & -\bar{x}_1\bar{x}_2 & ... & -\bar{x}_1\bar{x}_M \\ ... & ... & ... & ... \\ -\bar{x}_M\bar{x}_1 & -\bar{x}_M\bar{x}_2 & ... & -\bar{x}_M^2 \end{pmatrix} + \begin{pmatrix} \frac{1}{N}x_1^Tx_1 & \frac{1}{N}x_1^Tx_2 & ... & \frac{1}{N}x_1^Tx_M \\ ... & ... & ... & ... \\ \frac{1}{N}x_M^Tx_1 & \frac{1}{N}x_M^Tx_2 & ... & \frac{1}{N}x_M^Tx_M \end{pmatrix}$$

Ahora la primera matriz la podemos escribir como la multiplicación de dos vectores:

$$\begin{pmatrix} -\bar{x}_1^2 & -\bar{x}_1\bar{x}_2 & ... & -\bar{x}_1\bar{x}_M \\ ... & ... & ... & ... \\ -\bar{x}_M\bar{x}_1 & -\bar{x}_M\bar{x}_2 & ... & -\bar{x}_M^2 \end{pmatrix} = -\begin{pmatrix} \bar{x}_1 \\ ... \\ \bar{x}_M \end{pmatrix}\begin{pmatrix} \bar{x}_1 & ... & \bar{x}_M \end{pmatrix}$$

y en la segunda matriz hacer lo mismo sacando previamente $\frac{1}{N}$ factor común:

$$\begin{pmatrix} \frac{1}{N}x_1^Tx_1 & \frac{1}{N}x_1^Tx_2 & ... & \frac{1}{N}x_1^Tx_M \\ ... & ... & ... & ... \\ \frac{1}{N}x_M^Tx_1 & \frac{1}{N}x_M^Tx_2 & ... & \frac{1}{N}x_M^Tx_M \end{pmatrix} = \frac{1}{N}\begin{pmatrix} x_1^T \\ ... \\ x_M^T \end{pmatrix}\begin{pmatrix} x_1 & ... & x_M \end{pmatrix} = \frac{1}{N}X^TX$$

Y por tanto hemos llegado a que

$$cov(X) = -\begin{pmatrix} \bar{x}_1 \\ ... \\ \bar{x}_M \end{pmatrix}\begin{pmatrix} \bar{x}_1 & ... & \bar{x}_M \end{pmatrix} + \frac{1}{N}X^TX$$

Pregunta 2. Cosiderar la matriz hat definida en regresión, $H = X(X^TX)^{-1}X^T$, donde X es una matriz $N \times (d+1)$, y X^TX es invertible.

a) Mostrar que H es simétrica.

b) Mostrar que $H^K = H$ para cualquier entero $K > 0$.

a) H es simétrica si $H = H^T$. Veamos si esto se cumple. $H^T = (X(X^TX)^{-1}X^T)^T = (X^T)^T((X^TX)^{-1})^TX^T = X((X^TX)^{-1})^TX^T$ puesto que la traspuesta de una traspuesta es ella misma y por las propiedades de las matrices, $(AB)^T = B^TA^T$. Dado que H se diferencia de H^T sólo en la parte central, para que sean iguales tenemos que probar que $(X^TX)^{-1} = ((X^TX)^{-1})^T$. Como, de nuevo por las propiedades de las matrices, tenemos que $(A^{-1})^T = (A^T)^{-1}$ podemos desarrollar la parte de la derecha: $((X^TX)^{-1})^T = ((X^TX)^T)^{-1} = (X^T(X^T)^T)^{-1} = (X^TX)^{-1}$, que es la parte de la izquierda, luego $H^T = H$ y por tanto H es simétrica.

b) Tenemos que probar que $H = H^K$ para todo $K \in \mathbb{N}$. Lo vamos a hacer por inducción.
Para $K = 2$, que será nuestro caso base, tenemos que $H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)$. Si reagrupamos de la forma $X(X^TX)^{-1}(X^TX(X^TX)^{-1})X^T$ y utilizamos que $AA^{-1} = I$, donde I es la identidad de orden q con q es el número de filas de A, nos queda que $X(X^TX)^{-1}(X^TX(X^TX)^{-1})X^T = X(X^TX)^{-1}IX^T) = X(X^TX)^{-1}X^T = H$, con lo que lo probamos para $K = 2$.
Vamos ahora a suponerlo cierto para $K = n$, es decir, suponemos que $H^n = H$ y vamos a probarlo para $K = n + 1$:
$H^{n+1} = H^nH$ utilizamos la hipótesis de inducción y $H^nH = HH = H^2$, donde utilizamos el caso base y nos queda que $H^2 = H$ y por tanto $H^{n+1} = H$, como queríamos demostrar.

Pregunta 3. Resolver el siguiente problema: Encontrar el punto (x_0, y_0) sobre la línea $ax + by + d = 0$ que esté más cerca del punto (x_1, y_1).

Vamos a resolver este problema utilizando multiplicadores de Lagrange, como en el ejercicio pedido para el bonus 2, es decir, el ejercicio 3 del apartado de matrices y optimización, del que vamos a utilizar también la expresión de la distancia de dos curvas en el plano: $\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$, donde (x_0, y_0) está sobre una de las curvas y (x_1, y_1) está sobre la otra. En nuestro caso el punto (x_1, y_1) es un punto fijo en el plano y el punto (x_0, y_0) está sobre la línea $ax + by + d = 0$. Tenemos por tanto el siguiente problema:

$$\underset{(x_0,y_0)}{mín} \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2} \;\; Sujeto \;\; a \;\; ax_0 + by_0 + d = 0$$

Definimos la función langrangiana

$$\mathcal{L}(x_0, y_0, \lambda) = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2} - \lambda(ax_0 + by_0 + d)$$

y la solución a este problema es la solución del sistema de ecuaciones $\nabla_{(x_0,y_0,\lambda)}\mathcal{L}(x_0, y_0, \lambda) = 0$. Calculamos $\nabla_{(x_0,y_0,\lambda)}\mathcal{L}(x_0, y_0, \lambda)$:

$$\nabla_{(x_0,y_0,\lambda)}\mathcal{L}(x_0, y_0, \lambda) = (\frac{2(x_0 - x_1)}{2\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}} - \lambda a, \frac{2(y_0 - y_1)}{2\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}} - \lambda b, ax_0 - by_0 + d)$$

Si llamamos $A = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$ tenemos el siguiente sistema de 3 ecuaciones:

$$\frac{x_0 - x_1}{A} - \lambda a = 0$$

$$\frac{y_0 - y_1}{A} - \lambda b = 0$$

$$ax_0 + by_0 + d = 0$$

Ahora, de las dos primeras ecuaciones vamos a despejar λ y a igualar. De la primera ecuación,

$$\lambda = \frac{x_0 - x_1}{aA}$$

y de la segunda ecuación,

$$\lambda = \frac{y_0 - y_1}{bA}$$

Si igualamos,

$$\frac{x_0 - x_1}{aA} = \frac{y_0 - y_1}{bA} \Rightarrow a(y_0 - y_1) = b(x_0 - x_1) \Rightarrow x_0 = \frac{a}{b}(y_0 - y_1) + x_1$$

Ahora en la tercera ecuación vamos a sustituir x_0 por la expresión que acabamos de obtener:

$$ax_0 + by_0 + d = a(\frac{a}{b}(y_0 - y_1) + x_1) + by_0 + d = 0 \Rightarrow \frac{a^2}{b}y_0 - \frac{a^2}{b}y_1 + ax_1 + by_0 + d = 0$$

Despejamos y_0:

$$y_0(\frac{a^2}{b}+b) = -d - ax_1 + \frac{a^2}{b}y_1 \Rightarrow y_0 = \frac{-d - ax_1 + \frac{a^2}{b}y_1}{\frac{a^2+b^2}{b}} = \frac{-bd - abx_1 + a^2y_1}{a^2+b^2}$$

y con esto podemos obtener x_0 ya que los datos x_1 e y_1 son fijos (conocidos):

$$x_0 = \frac{a}{b}(y_0 - y_1) + x_1 = \frac{a}{b}(\frac{-bd - abx_1 + a^2y_1}{a^2+b^2} - y_1) + x_1 = \frac{-ad - a^2x_1 + \frac{a^3}{b}y_1}{a^2+b^2} - \frac{a}{b}y_1 + x_1 =$$

$$= \frac{-ad}{a^2+b^2} + \frac{-a^2x_1}{a^2+b^2} + x_1 + \frac{a^3y_1}{b(a^2+b^2)} - \frac{a}{b}y_1 = \frac{-ad}{a^2+b^2} + \frac{-a^2x_1 + a^2x_1 + b^2x_1}{a^2+b^2} + \frac{a^3y_1 - a^3y_1 - ab^2y_1}{b(a^2+b^2)} =$$

$$= \frac{-ad}{a^2+b^2} + \frac{b^2x_1}{a^2+b^2} + \frac{-aby_1}{a^2+b^2} = \frac{-ad + b^2x_1 - aby_1}{a^2+b^2}$$

Con lo que el punto de la recta $ax_0 + by_0 + d = 0$ más cercano al punto (x_1, y_1) es

$$(x_0, y_0) = (\frac{b^2x_1 - aby_1 - ad}{a^2+b^2}, \frac{a^2y_1 - abx_1 - bd}{a^2+b^2})$$

Pregunta 4. Consideremos el problema de optimización lineal con restricciones definido por

$$\underset{\mathbf{z}}{Min} \;\; \mathbf{c}^T\mathbf{z}$$

$$Sujeto\ a\ A\mathbf{z} \leq b$$

donde **c** y **b** son vectores y A es una matriz.

a) Para un conjunto de datos linealmente separable mostrar que para algún **w** se debe verificar la condición $y_n\mathbf{w}^T\mathbf{x}_n > 0$ para todo $(\mathbf{x}_n, y_n)$ del conjunto.

b) Formular un problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, identifique quiénes son A, **z**, **b** y **c** para este caso.

a) Si el conjunto de datos es linealmente separable entonces existe un **w** que deja los datos con etiqueta positiva a un lado y los de etiqueta negativa a otro lado, es decir, que clasifica bien todos los datos, y entonces $signo(y_n) = signo(\mathbf{w}^T\mathbf{x}_n)$ para cada n puesto que todos están bien clasificados y por tanto $y_n\mathbf{w}^T\mathbf{x}_n > 0$.

b) Vamos a suponer como en el apartado anterior que los datos son linealmente separables y que tenemos N datos. Entonces como hemos visto los datos que estén bien clasificados tendrán $y_n\mathbf{w}^T\mathbf{x}_n > 0$ y los que estén mal clasificados tendrán $y_n\mathbf{w}^T\mathbf{x}_n < 0$. Nosotros queremos que $\sum_{i=1}^N y_i\mathbf{w}^T\mathbf{x}_n$ sea lo mayor posible para que todos los datos estén bien clasificados y con la mayor distancia a **w**, luego queremos minimizar $-\sum_{i=1}^N y_i\mathbf{w}^T\mathbf{x}_n$ y por tanto $\mathbf{c}^T = (-1, ..., -1)$ de dimensión N y $\mathbf{z} = (y_1\mathbf{w}^T\mathbf{x}_1, ..., y_N\mathbf{w}^T\mathbf{x}_N)$ sujetos a que cada $y_i\mathbf{w}^T\mathbf{x}_n > 0$ para obligar a que todos los puntos estén bien clasificados, de forma que

$$A = \begin{pmatrix} -1 & 0 & ... & 0 \\ 0 & -1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & -1 \end{pmatrix}$$

puesto que la restricción que tenemos en el enunciado es $\leq$ y le tenemos que dar la vuelta y $b^T = (0, ..., 0)$ de dimensión N.

Pregunta 5. Probar que en el caso general de funciones con ruido se verifica que $\mathbb{E}_{\mathcal{D}}[E_{out}] = \sigma^2 + bias + var$ (ver transparencias de clase).

Tenemos que calcular $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})]$ teniendo en cuenta que nuestra función f la recibimos con ruido, es decir, que tenemos $y_n = f(x_n) + \varepsilon$, donde ε tiene media cero y varianza σ^2. Sabemos que entonces $E_{out}(h) = \mathbb{E}_x[(h(x) - f(x) - \varepsilon)^2]$. Para $h = g^{\mathcal{D}}$ que es lo que tenemos que calcular, $E_{out}(g^{\mathcal{D}}) = \mathbb{E}_x[(g^{\mathcal{D}}(x) - f(x) - \varepsilon)^2]$. Al introducir ε, lo que tenemos que calcular $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})]$ se convierte en $\mathbb{E}_{\mathcal{D},\varepsilon}[E_{out}(g^{\mathcal{D}})]$. Vamos a desarrollarlo:

$$\mathbb{E}_{\mathcal{D},\varepsilon}[E_{out}(g^{\mathcal{D}})] = \mathbb{E}_{\mathcal{D},\varepsilon}[\mathbb{E}_x[(g^{\mathcal{D}}(x) - f(x) - \varepsilon)^2]] = \mathbb{E}_x[\mathbb{E}_{\mathcal{D},\varepsilon}[(g^{\mathcal{D}}(x) - f(x) - \varepsilon)^2]] =$$

Desarrollamos el cuadrado agrupando en un término $(f(x) + \varepsilon)$:

$$= \mathbb{E}_x[\mathbb{E}_{\mathcal{D},\varepsilon}[g^{\mathcal{D}}(x)^2 + (f(x) + \varepsilon)^2 - 2g^{\mathcal{D}}(x)(f(x) + \varepsilon)]] =$$

Desarrollamos ahora el cuadrado que queda y utilizamos que f y ε son independientes de $\mathcal{D}$ y por tanto los podemos sacar de la esperanza con respecto a $\mathcal{D}$, al igual que g es independiente de ε.

$$= \mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] + f(x)^2 + \mathbb{E}_\varepsilon[\varepsilon^2 + 2f(x)\varepsilon] - 2\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]\mathbb{E}_\varepsilon[(f(x) + \varepsilon)]] =$$

Utilizamos la definición de $\bar{g} = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]$ y separamos las esperanzas que quedan:

$$\mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] + f(x)^2 + \mathbb{E}_\varepsilon[\varepsilon^2] + \mathbb{E}_\varepsilon[2f(x)\varepsilon] - 2\bar{g}(x)f(x) - 2\bar{g}(x)\mathbb{E}_\varepsilon[\varepsilon]] =$$

Ahora, como ε es independiente de f y de x, $\mathbb{E}_\varepsilon[2f(x)\varepsilon] = 2f(x)\mathbb{E}_\varepsilon[\varepsilon]$ y como hemos dicho previamente, la media de ε es 0, luego nos queda:

$$= \mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] + f(x)^2 + \mathbb{E}_\varepsilon[\varepsilon^2] - 2\bar{g}(x)f(x)] =$$

Sumamos y restamos $\bar{g}(x)$:

$$= \mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] - \bar{g}(x) + \bar{g}(x) - 2\bar{g}(x)f(x) + f(x)^2 + \mathbb{E}_\varepsilon[\varepsilon^2]]$$

Veamos que $\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] - \bar{g}(x) = \mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2]$. Partimos de la derecha y desarrollamos el cuadrado: $\mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2] = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2 + \bar{g}(x)^2 - 2\bar{g}(x)g^{\mathcal{D}}(x)] = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] + \bar{g}(x)^2 - 2\bar{g}(x)\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)] = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] + \bar{g}(x)^2 - 2\bar{g}(x)\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)^2] - \bar{g}(x)^2$, como queríamos probar.
Ahora, juntando esto con que $\bar{g}(x) - 2\bar{g}(x)f(x) + f(x)^2 = (\bar{g}(x) - f(x))^2$ y que $\mathbb{E}_\varepsilon[\varepsilon^2]$ es la varianza de ε porque ε tiene media 0 (por definición, la varianza es la media de $(\varepsilon - media)^2$) tenemos $\mathbb{E}_\varepsilon[\varepsilon^2] = \sigma^2$ la expresión anterior nos queda:

$$\mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2] + (\bar{g}(x) - f(x))^2 + \sigma^2] = \mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2]] + \mathbb{E}_x[(\bar{g}(x) - f(x))^2] + \sigma^2 =$$

Y sabemos que $\mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2] = var(x)$ y $(\bar{g}(x) - f(x))^2 = bias(x)$, luego nos queda:

$$= \mathbb{E}_x[var(x)] + \mathbb{E}_x[bias(x)] + \sigma^2 = var + bias + \sigma^2$$

que era lo que nos pedía el ejercicio.

Pregunta 6. Consideremos las mismas condiciones generales del enunciado del Ejercicio 2 del apartado de Regresión de la relación de ejercicios 2. Considerar ahora $\sigma = 0{,}1$ y $d = 8$, ¿cuál es el más pequeño tamaño muestral que resultará en un valor esperado de E_{in} mayor de 0,008?

En dicho ejercicio nos dice que el error esperado de entrenamiento es $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2(1 - \frac{d+1}{N})$. Sustituyendo con los datos que nos da este ejercicio tenemos $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = (0{,}1)^2(1-\frac{8+1}{N})$ y queremos obtener N para que $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] > 0{,}008$, es decir, $(0{,}1)^2(1-\frac{8+1}{N}) > 0{,}008$.

$$(0{,}1)^2(1 - \frac{8+1}{N}) = 0{,}01 - \frac{0{,}09}{N} > 0{,}008 \Rightarrow -\frac{0{,}09}{N} > -0{,}002 \Rightarrow \frac{0{,}09}{N} < 0{,}002$$

$$\Rightarrow N > \frac{0{,}09}{0{,}002} \Rightarrow N > 45$$

Por tanto, el más pequeño muestral para que el valor esperado de E_{in} sea mayor de 0,008 es 46.

Pregunta 7. En regresión logística mostrar que

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

Argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.

Derivamos con respecto al vector $\mathbf{w}$.

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

Multiplicamos arriba y abajo de la fracción por $e^{y_n\mathbf{w}^T\mathbf{x}_n}$ y nos queda:

$$-\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$$

que era el primer sumando a probar. Para el segundo, volvemos a la primera expresión que hemos obtenido y recordamos que $\sigma(s) = \frac{e^s}{1+e^s}$, con lo que tomando como $s = -y_n\mathbf{w}^T\mathbf{x}_n$ nos queda

$$-\frac{1}{N}\sum_{n=1}^{N} y_n\mathbf{x}_n\sigma(-y_n\mathbf{w}^T\mathbf{x}_n)$$

En cuanto a que un ejemplo mal clasificado contribuye más al gradiente que uno bien clasificado, vamos a fijarnos en la expresión $\nabla E_{in} = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$. Si está mal clasificado, entonces $signo(y_n) \neq signo(\mathbf{w}^T\mathbf{x}_n)$ y por tanto $y_n\mathbf{w}^T\mathbf{x}_n$ será negativo y por tanto $e^{y_n\mathbf{w}^T\mathbf{x}_n}$ será más pequeño que si fuera positivo, es decir, que si estuviera bien clasificado al ser la exponencial una función creciente. Al ser más pequeño el denominador, la fracción es más grande, y por tanto contribuye más al gradiente.

Pregunta 8. Definimos el error en un punto $(\mathbf{x}_n, y_n)$ por

$$\mathbf{e}_n(\mathbf{w}) = max(0, -y_n \mathbf{w}^T \mathbf{x}_n)$$

Argumentar que el algoritmo PLA puede interpretarse como SGD sobre $\mathbf{e}_n$ con tasa de aprendizaje $\eta = 1$.

La regla de actualización del SGD es $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \mathbf{e}_n(\mathbf{w})$ y la del PLA es $\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n$ cuando $\mathbf{w}$ está clasificando mal $\mathbf{x}_n$ y $\mathbf{w} \leftarrow \mathbf{w}$ si lo está clasificando bien. Veamos si podemos llegar a esto partiendo del SGD. Como $\eta = 1$ nos queda la regla $\mathbf{w} \leftarrow \mathbf{w} - \nabla \mathbf{e}_n(\mathbf{w})$. Vamos a calcular $\nabla \mathbf{e}_n(\mathbf{w})$ con $\mathbf{e}_n(\mathbf{w}) = max(0, -y_n \mathbf{w}^T \mathbf{x}_n)$, que será en cada caso la derivada de 0 si éste es mayor o la derivada de $-y_n \mathbf{w}^T \mathbf{x}_n$ si éste es el mayor:

$$\nabla \mathbf{e}_n(\mathbf{w}) = \begin{cases} 0 & si \quad 0 \geq -y_n \mathbf{w}^T \mathbf{x}_n \\ -y_n \mathbf{x_n} & si \quad 0 < -y_n \mathbf{w}^T \mathbf{x}_n \end{cases} \tag{1}$$

Ahora, $0 \geq -y_n \mathbf{w}^T \mathbf{x}_n$ si $y_n \mathbf{w}^T \mathbf{x}_n \geq 0$ y $0 < -y_n \mathbf{w}^T \mathbf{x}_n$ si $y_n \mathbf{w}^T \mathbf{x}_n < 0$. Introduciendo esto en la regla de actualización del SGD, nos queda que $\mathbf{w} \leftarrow \mathbf{w} - 0$ si $y_n \mathbf{w}^T \mathbf{x}_n \geq 0$, es decir, si el dato está bien clasificado, y $\mathbf{w} \leftarrow \mathbf{w} - (-y_n \mathbf{x_n})$ si $0 < -y_n \mathbf{w}^T \mathbf{x}_n$, es decir, si está mal clasificado, que es justo la regla de actualización del PLA, como queríamos probar.

Pregunta 9. El ruido determinista depende de $\mathcal{H}$, ya que algunos modelos aproximan mejor f que otros.

a) Suponer que $\mathcal{H}$ es fija y que incrementamos la complejidad de f.

b) Suponer que f es fija y decrementamos la complejidad de $\mathcal{H}$.

Contestar para ambos escenarios: ¿En general subirá o bajará el ruido determinista? ¿La tendencia a sobreajustar será mayor o menor? (Ayuda: analizar los detalles que influencian el sobreajuste).

a) En este caso aumentará el ruido determinista, ya que el área en el que se diferencian f y h (donde h es el mejor ajuste de f en $\mathcal{H}$) aumentará en tanto que f va aumentando complejidad y $\mathcal{H}$ es fijo (no podemos ajustar con un polinomio de más grado del que nos deje $\mathcal{H}$ cuando f lo requiere) y por tanto según aumente la complejidad de f ajustaremos cada vez peor.
La tendencia a sobreajustar será menor ya que según aumente la complejidad de la f estaremos ajustando peor con h y por tanto no se podrán sobreajustar los datos.

b) Dependerá de la complejidad previa que tuviera $\mathcal{H}$. Aumentará el ruido determinista cuando la complejidad de $\mathcal{H}$ sea demasiado baja, ya que de nuevo ajustaremos f cada vez peor. Sin embargo si H era muy alta y estábamos sobreajustando f, el ruido determinista irá bajando hasta que se obtenga el mejor ajuste posible y volverá a subir si seguimos bajando la complejidad de $\mathcal{H}$, llegando a lo comentado anteriormente.
La tendencia a sobreajustar será menor si estábamos sobreajustando los datos al tener muchos grados de libertad en $\mathcal{H}$, ya que al ir bajando la complejidad irá bajando el sobreajuste al ir ajustando cada vez peor.

Pregunta 10. La técnica de regularización de Tikhonov es bastante general al usar la condición

$$\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} \leq C$$

que define relaciones entre las w_i (la matriz Γ_i se denomina regularizados de Tikhonov)

a) Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$

b) Calcular Γ cuando $(\sum_{q=0}^{Q} w_q)^2 \leq C$

Argumentar si el estudio de los regulizadores de Tikhonov puede hacerse a través de las propiedades algebraicas de las matrices Γ.

a) Si queremos que $\sum_{q=0}^{Q} w_q^2 \leq C$ se tiene que dar que $\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} = \sum_{q=0}^{Q} w_q^2$, pero ya tenemos que $\mathbf{w}^T\mathbf{w} = \sum_{q=0}^{Q} w_q^2$, con lo que $\Gamma^T\Gamma$ tiene que ser la identidad matricial de orden el número de filas de Γ^T si queremos que esto se siga cumpliendo, es decir, que nos vale cualquier Γ invertible tal que $\Gamma^{-1} = \Gamma^T$ (ya que si Γ es invertible, $\Gamma^{-1}\Gamma = I$), es decir, nos vale cualquier matriz ortogonal.

b) Ahora tiene que darse que $\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} = (\sum_{q=0}^{Q} w_q)^2$, es decir, $\mathbf{w}^T\Gamma^T = \Gamma\mathbf{w} = (\sum_{q=0}^{Q} w_q, 0, 0)$ o su traspuesto, y esto lo tenemos si

$$\Gamma = \begin{pmatrix} 1 & 1 & ... & 1 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \end{pmatrix}$$

Veámoslo por ejemplo para $Q = 2$, donde tenemos 3 componentes de $\mathbf{w}$:

$$\mathbf{w}^T\Gamma^T = \begin{pmatrix} w_0 & w_1 & w_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} = (w_0 + w_1 + w_2, 0, 0)$$

$$\Gamma\mathbf{w} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} w_0 + w_1 + w_2 \\ 0 \\ 0 \end{pmatrix}$$

y entonces

$$\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} = \begin{pmatrix} w_0 + w_1 + w_2 & 0 & 0 \end{pmatrix} \begin{pmatrix} w_0 + w_1 + w_2 \\ 0 \\ 0 \end{pmatrix} = (w_0 + w_1 + w_2)^2$$

que era lo que queríamos.

En cuanto a si el estudio de los regulizadores de Tikhonov puede hacerse a través de las propiedades algebraicas de las matrices Γ, no creo que se pueda hacer de forma generalizada sólo con las propiedades algebraicas de las matrices Γ, ya que hay casos en los que sí se pueden utilizar dichas propiedades (como en el caso a), donde nos vale una Γ ortogonal) y casos en los que no, como el b), en el que Γ no tiene por qué tener ninguna propiedad algebraica.

2. Bonus

Pregunta 11. Considerar la matriz hat $H = X(X^TX)^{-1}X^T$. Sea X una matriz $N \times (d+1)$, y X^TX invertible. Mostrar que traza(H)=$d+1$, donde traza significa la suma de los elementos de la diagonal principal.

Vamos a utilizar la siguiente propiedad de las matrices: si A es una matriz $P \times Q$ y B es una matriz $Q \times P$, se tiene que traza(AB) = traza(BA). En nuestro caso tenemos que traza(H) = traza($X(X^TX)^{-1}X^T$) y $X(X^TX)^{-1}$ es una matriz $N \times (d+1)$ (ya que $(X^TX)^{-1}$ es una matriz cuadrada de orden $d+1$) y X^T es una matriz $(d+1) \times N$, con lo que podemos hacer traza($X(X^TX)^{-1}X^T$) = traza($(X^TX)^{-1}X^TX$) = traza(I), donde I es la identidad de orden $d+1$ puesto que tanto $(X^TX)^{-1}$ como X^TX son matrices cuadradas de orden $d+1$ por ser X una matriz $N \times (d+1)$. Finalmente, traza(I) = $d+1$, ya que es la suma de $d+1$ unos, lo que nos da lo que queríamos probar.

3. Bibliografía

1. Las transparencias de clase disponibles en decsai.
2. El libro *Learning from data*, de Abu-Mostafa, Magdon-Ismail & Lin