Cuestionario 2

Luis Suárez Lloréns

Cuestión 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.

Respuesta. Siendo X la matriz, supongamos que $\bar{X}$ una matriz que tiene como valor de las columnas la media de las columnas de X. Entonces:

$$cov(X) = \frac{1}{N}\left(X - \bar{X}\right)^T\left(X - \bar{X}\right)$$

Pues con esta operación, en la posición (i, j) de $cov(X)$ obtenemos:

$$\frac{1}{N}\sum_{k=1}^{N}(x_{k,i} - \overline{x_i})(x_{k,j} - \overline{xj})$$

siendo $\overline{x_i}$ la media de la columna i.

Falta definir entonces $\bar{X}$:

$$\bar{X} = \frac{1}{N}1_{nxn}X$$

Siendo 1_{nxn} la matriz con todos sus valores 1, y n es el número de datos que tenemos por variable.

Cuestión 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.

Respuesta. a)

$$H^T = \left(X\left(X^TX\right)^{-1}X^T\right)^T = X^{T^T}\left(X^TX\right)^{-1^T}X^T =$$

$$X\left(X^TX^{T^T}\right)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

b) Si $H^2 = H$, entonces por inducción, tendríamos que $H^K = H$. Por tanto, sólo vamos a demostrar $H^2 = H$.

$$H^2 = HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T =$$
$$X(X^TX)^{-1}\left(X^TX(X^TX)^{-1}\right)X^T = X(X^TX)^{-1}X^T = H$$

Cuestión 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).

Respuesta. Vamos a resolver este problema, utilizando multiplicadores de Lagrange. Los multiplicadores de Lagrange tratan de buscar el máximo de una función. En nuestro caso, buscamos el mínimo de una función, que es lo mismo que buscar el máximo de menos dicha función. Además, no necesitamos la raiz a la hora de calcular la distancia, pues el punto que obtengamos va a ser el mismo. Así pues, obtenemos:

$$\mathcal{L}(x, y, \lambda) = -\left((x_1 - x)^2(y_1 - y)^2\right) - \lambda(ax + by + c)$$

$$\nabla\mathcal{L}(x, y, \lambda) =< 2(x_1 - x) - a\lambda, 2(y_1 - y) - a\lambda, ax + by + c >$$

Ahora, igualamos $\nabla\mathcal{L}(x, y, \lambda)$ a 0 y resolvemos el sistema de ecuaciones:

$$\begin{cases} 2(x_1 - x) - a\lambda = 0 \\ 2(y_1 - y) - a\lambda = 0 \\ ax + by + c = 0 \end{cases}$$

De las dos primeras, podemos obtener expresiones de x e y en función de λ:

$$2x_1 - a\lambda = 2x \Rightarrow x_1 - \frac{a\lambda}{2} = x$$

$$2y_1 - b\lambda = 2y \Rightarrow y_1 - \frac{b\lambda}{2} = y$$

Por último, para calcular λ, sustituimos en la última ecuación del sistema y resolvemos:

$$ax + by + c = 0 \Rightarrow a\left(x_1 - \frac{a\lambda}{2}\right) + b\left(y_1 - \frac{b\lambda}{2}\right) + c = 0$$

$$\Rightarrow ax_1 + by_1 + c = \left(\frac{a^2}{2} + \frac{b^2}{2}\right)\lambda \Rightarrow \lambda = \frac{ax_1 + by_1 + c}{\left(\frac{a^2}{2} + \frac{b^2}{2}\right)}$$

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

$$\text{Min}_{\mathbf{z}}\mathbf{c}^T\mathbf{z}$$

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

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.

Respuesta. a) Dado que los datos son linealmente separables, existe un w que separe perfectamente los datos. Eso significa que el signo de $w^T x_n$ es igual al signo de y_n para todo n. Entonces, si $y_n > 0$, $w^T x_n$ también lo será, luego el producto será positivo. Si $y_n < 0$, $w^T x_n$ sería negativo, y el producto saldría también positivo. Luego tenemos que:

$$y_n w^T x_n > 0$$

b) Para transformar el problema del apartado a, en el problema que tenemos de programación lineal, tenemos que llevar la condición del apartado a, es decir, $y_n w^T x_n > 0$ a la condición $Az \leq b$.

El primer paso es tener el signo de comparación correcto, esto lo conseguimos multiplicando por -1 la primera desigualdad, obteniendo $-y_n w^T x_n \leq 0$.

Tras esto, tenemos que $-y_n w^T x_n \leq 0$ es cierto para cada dato. Luego tenemos que transformar las condiciones que obtenemos —una por cada dato— a una expresión matricial de la forma $Az \leq b$. Es claro que z será nuestro vector de pesos w y que b debe de ser el vector 0.

Nos queda por tanto la definición de A. Para cada dato de tamaño N, la operación que realizamos en la búsqueda del hiperplano separador es:

$$-yw^T x = \sum_{i=1}^{N} -yw_i x_i = \sum_{i=1}^{N} (-yx_i) w_i$$

Luego son esos valores $(-yx_i)$ los que tenemos que almacenar en nuestra matriz A para poder realizar la operación. Por tanto, la matriz A será una matriz de N columnas por d —número de datos— filas, donde cada fila se define como $(-yx_i)$ para un dato distinto del conjunto.

Por último, notar que no hemos dicho nada de c. En realidad, la expresión $Az \leq b$ anterior ya contiene todas las restricciones del problema. Luego podríamos usar como c el vector 0, para no imponer ninguna restricción adicional.

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

Respuesta. La función que vamos a aproximar, ahora tiene ruido. Es decir, tenemos una función $F(x) = f(x) + \epsilon$ con ϵ un ruido de una normal de media 0 y desviación típica σ.

Repetimos los pasos de la transparencia, pero utilizando la función $F(x)$ y llegamos a:

$$E_D[E_o ut(g^{(D)})] = E_x[E_D(g^D(x)^2) - \tilde{g}(x)^2 + \tilde{g}(x)^2 - 2\tilde{g}(x)F(x) + F(x)^2 =$$

$$E_x[\text{var}(X) + \tilde{g}(x)^2 - 2\tilde{g}(x)f(x) + f(x)^2 - 2\tilde{g}(x)\epsilon + 2f(x)\epsilon + \epsilon^2] =$$
$$E_x[\text{var}(X) + \text{bias}(x) + 2(f(x) - \tilde{g}(x))\epsilon + \epsilon^2]$$

Pero la esperanza de ϵ es 0, pues es la media de su distribución. La esperanza de ϵ^2 es la varianza de su distribución. Luego nos queda:

$$E_D[E_{out}(g^{(D)})] = \text{bias} + \text{var} + \sigma^2$$

Cuestión 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?

Respuesta. Partimos de la ecuación del ejercicio 2, y asignamos los valores que nos dice el ejercicio:

$$E_D[E_{in}(w_{lin})] = \sigma^2\left(1 - \frac{d+1}{N}\right) = 0.01 * \left(1 - \frac{9}{N}\right)$$

Ahora igualamos con 0.008 y calculamos N:

$$0.01 * \left(1 - \frac{9}{N}\right) = 0.008 \Rightarrow -\frac{9}{N} = \frac{0.008}{0.01} - 1 = -0.2 \Rightarrow \frac{9}{0.2} = 45 = N$$

Entonces, con $N = 45$, obtenemos justo un valor esperado de E_{in} de 0.008. Luego para obtener un valor esperado mayor de 0.008, necesitamos $45 < N$. El mínimo valor que cumple esto es $N = 46$.

Cuestión 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.

Respuesta. Primero, destacar que:

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

Luego de las dos expresiones que tenemos, aplicando la segunda expresión de $\sigma(x)$ a la segunda de ∇E_{in} obtenemos directamente la primera expresión de ∇E_{in}.

Por tanto, nos queda calcular directamente la expresión de ∇E_{in} y ver que es igual que cualquiera de las dos:

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N} \log 1 + e^{-y_n w^T x_n}$$

$$\nabla E_{in} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1 + e^{-y_n w^T x_n}}\left(1 + e^{-y_n w^T x_n}\right)' =$$

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

Sustituimos por la primera expresión de $\sigma(-y_n w^T x_n)$ y obtenemos:

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

Por último, si un dato está bien clasificado, $y_n w^T x_n$ es positivo, y si lo clasifica mal, el número es negativo. Si observamos la primera expresión de ∇E_{in}, vemos que, si $y_n w^T x_n$ es positivo, dividimos $y_n x_n$ por un número mayor que cuando es negativo.

Por tanto, si un dato está mal clasificado, tendrá más importancia —pues dividimos por un número más pequeño que cuando está bien clasificado—.

Cuestión 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 $\nu = 1$.

Respuesta. Para empezar, el algoritmo PLA actualiza sus pesos en función de un único dato. Por tanto, si fuera un SGD, el tamaño de "mini-batch" sería de 1.

La función de actualización del PLA es la siguiente:

- Si está bien clasificado, no se actualiza.
- Si está mal clasificado, se actualiza usando $y_n w^T x_n$.

En el caso del SGD presentado en el ejercicio, si un dato está bien clasificado, $-y_n w^T x_n$ es negativo y por tanto, el máximo es 0. Es decir, no se actualiza.

En el caso de un dato mal clasificado, $-y_n w^T x_n$ es positivo. Sólo que da por ver la diferencia de signo, pero en el caso del SGD se resta el valor dado por $\mathbf{e}_n$, luego el valor es el mismo, $y_n w^T x_n$.

Por último, al ser el mismo valor exactamente, ν tiene que ser 1.

Cuestión 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).

Respuesta. a) Al aumentar la complejidad de la función objetivo, aumentamos el ruido determinista. Esto se debe a que la mejor aproximación de la clase $\mathcal{H}$ no puede modelar a la función objetivo f.

Como hemos dicho, se aumenta el ruido, dando como sabemos un mayor sobreajuste, pues la función obtenida intentará aprender los valores obtenidos, que tienen errores, y no la mejor función posible de $\mathcal{H}$.

b) También se aumenta el ruido determinista, pues al reducirse la complejidad de $\mathcal{H}$ disminuye, y por tanto la mejor función de $\mathcal{H}$ representa cada vez peor la función real.

Ahora bien, al tener un modelo más sencillo, la función que obtenemos de $\mathcal{H}$ sólo obtiene las características generales de la función real. Esto hace que, pese a tener más error determinista, la función obtenida tiene más capacidad de generalización.

Visto desde otra perspectiva, ir hacia un modelo más sencillo, aumenta un poco el bias, pero disminuye mucho la varianza, consiguiendo un menor error y un menor sobreajuste.

Cuestión 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 $(\sum_{q=0}^{Q} w_q)^q \leq C$

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

Respuesta. Siendo w, un vector columna de n elementos:

- **a)** La matriz que tenemos que seleccionar es la identidad I_{nxn}.
- **b)** La matriz que tenemos que seleccionar es la matriz cuadrada nxn que tiene la primera fila rellena de 1 y el resto 0.

En cuanto al estudio de los regularizadores de Tikhonov, es claro que el único factor que puede cambiar son las matrices Γ. Por tanto, su estudio y propiedades depende directamente de las propiedades de la matriz que seleccionamos como Γ.

Bonus

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

Respuesta. Vamos a utilizar la propiedad de que la traza de AB es igual a la traza de BA. Usando dicha propiedad, calculamos la traza de H. Destacar que la matriz H es cuadrada y de dimensión $d+1$ ($d+1=N$).

$$\text{traza}(H) = \text{traza}(X(X^TX)^{-1}X^T) = \text{traza}(X^TX(X^TX)^{-1}) = \text{traza}(I) = d+1$$

Siendo I la matriz identidad de tamaño $d+1$.