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

$$Min_{\mathbf{z}}\mathbf{c}^T\mathbf{z}$$

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

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.

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).

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?

Pregunta 7. En regresión logística mostrar que

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

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

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 $\nu = 1$.

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).

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

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

2. Bonus

Pregunta 11. 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.

3. Bibliografía