Aprendizaje Automático: Cuestionario 2

Anabel Gómez Ríos

14 de mayo de 2016

1. Cuestiones

Pregunta 1. Sean $\mathbf{x}$ e $\mathbf{y}$ 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 $\mathbf{z}$. 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:

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

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:

$$\begin{pmatrix} \frac{1}{N}x_1^T x_1 & \frac{1}{N}x_1^T x_2 & ... & \frac{1}{N}x_1^T x_M \\ ... & ... & ... & ... \\ \frac{1}{N}x_M^T x_1 & \frac{1}{N}x_M^T x_2 & ... & \frac{1}{N}x_M^T x_M \end{pmatrix} = \frac{1}{N}\begin{pmatrix} x_1^T \\ ... \\ x_M^T \end{pmatrix} \begin{pmatrix} x_1 & ... & x_M \end{pmatrix} = \frac{1}{N}X^T X$$

Y por tanto hemos llegado a que

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

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

a) Mostrar que H es simétrica.

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

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).

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 $\mathbf{w}$ 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 $H = X(X^TX)^{-1}X^T$. 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.

3. Bibliografía