Aprendizaje Automático (2015-2016)
Grado en Ingeniería Informática
Universidad de Granada

Cuestionario 2

Laura Tirado López

13 de mayo de 2016

1. Sean x e y dos vectores de observaciones de tamaño N. Sea $cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$ la covarianza de dichos vectores, donde $\bar{z}$ 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.

   La matrix X sería:

$$X = \begin{pmatrix} x_1 & x_1' & \dots & x_1^m \\ x_2 & x_2' & \dots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_n' & \dots & x_n^m \end{pmatrix}$$

   formada por m columnas, siendo cada columna un vector de observaciones, y n filas (número de componentes de cada vector). Definimos v_i a la columna i de la matriz X. De esta forma la matriz $cov(X)$ sería:

$$cov(X) = \begin{pmatrix} cov(v_1,v_1) & cov(v_1,v_2) & \dots & cov(v_1,v_m) \\ cov(v_2,v_1) & cov(v_2,v_2) & \dots & cov(v_2,v_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(v_m,v_1) & cov(v_m,v_2) & \dots & cov(v_m,v_m) \end{pmatrix}$$

   siendo $cov(x,y)$ la definida en el enunciado.

   Las matrices X y $cov(X)$ pueden ser expresadas de la siguiente forma:

$$X := (x_{ij})_{m\times n}$$
$$cov(X) := (c_{ij})_{m\times m}$$

   donde

$$c_{ij} = \frac{1}{n}\sum_{i=1}^{n}(a_{ii} - \bar{a})(b_{ij} - \bar{b})$$

   teniendo en cuenta que $a_{ii}, b_{ij} \in X$.

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

   a) Mostrar que H es simétrica. Para demostrar que H es simétrica tenemos que demostrar que $\mathrm{H} = \mathrm{H}^{\mathrm{T}}$.

$$\mathrm{H}^{\mathrm{T}} = (\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}})^{\mathrm{T}}$$

La traspuesta de un producto es el producto de las traspuesta invertido por lo que:

$$(\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}})^{\mathrm{T}} = (\mathrm{X}^{\mathrm{T}})^{\mathrm{T}}((\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1})^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} = \mathrm{X}((\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1})^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}$$

Como $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es simétrica sabemos que su inversa también lo es, lo que significa que $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$. Teniendo en cuenta esto:

$$\mathrm{X}((\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1})^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} = H$$

Por tanto hemos demostrado que H es simétrica.

b) Mostrar que $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero positivo K. Para demostrarlo, basta con probar que H es idempotente, es decir, HH = H.

$$HH = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} \cdot \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} =$$

$$= \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{X})(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{I}\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} = \mathrm{H}$$

Como H es idempotente, $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero K.

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

El objetivo es minimizar la función $\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$ teniendo en cuenta que $ax_0 + bx_0 + d = 0$. Lo resolvemos como cualquier problema de optimización. Primero despejamos una de las coordenadas del punto:

$$y_0 = \frac{-d - ax_0}{b}$$

Sustituimos en la función a optimizar:

$$f = \sqrt{(x_1 - x_0)^2 + (y_1 - (\frac{-d - ax_0}{b}))^2} = \sqrt{(x_1 - x_0)^2 + (y_1 + \frac{d + ax_0}{b})^2}$$

Derivamos la función:

$$f' = \frac{a(ax_0 + d + y_1 b) - b^2(x_1 - x_0)}{b^2\sqrt{\frac{(ax_0+d+y_1b)^2+b^2(x_1-x_0)^2}{b^2}}}$$

Igualamos la derivada a 0 y despejamos x_0.

$$f' = 0$$

$$\frac{a(ax_0 + d + y_1 b) - b^2(x_1 - x_0)}{b^2\sqrt{\frac{(ax_0+d+y_1b)^2+b^2(x_1-x_0)^2}{b^2}}} = 0$$

$$a(ax_0 + d + y_1 b) - b^2(x_1 - x_0) = 0$$
$$a^2 x_0 + ad + ay_1 b - b^2 x_1 + b^2 x_0 = 0$$
$$x_0(a^2 + b^2) - b^2 x_1 + ad + ay_1 b = 0$$
$$x_0(a^2 + b^2) = b^2 x_1 - ad - ay_1 b$$
$$x_0 = \frac{b^2 x_1 - ad + ay_1 b}{a^2 + b^2}$$

Por tanto:

$$y_0 = \frac{-d - a(\frac{b^2 x_1 - ad + ay_1 b}{a^2 + b^2})}{b} = \frac{-2da^2 + a^2 b y_1 + ab^2 x_1 - db^2}{b(a^2 + b^2)}$$

El punto (x_0, y_0) es

$$(\frac{b^2 x_1 - ad + ay_1 b}{a^2 + b^2}, \frac{-2da^2 + a^2 b y_1 + ab^2 x_1 - db^2}{b(a^2 + b^2)})$$

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

$$Min_z \mathrm{c^T z}$$
$$\text{Sujeto a } \mathrm{Az} \leq \mathbf{b}$$

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

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

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

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

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 $d = 8$, cual es el más pequeño tamaño muestral que resultará en un valor esperado de mayor de 0,008?. Para ver el tamaño muestral más pequeño, sustituimos los datos en la fórmula y despejamos N:

$$\mathbb{E}[E_{in}(w_{lin})] = \sigma^2(1 - \frac{d+1}{N})$$
$$0{,}008 = 0{,}01(1 - \frac{8+1}{N})$$

$$0{,}008 = 0{,}01(1 - \frac{9}{N})$$
$$0{,}008 = 0{,}01 - \frac{0{,}09}{N}$$
$$0{,}008 - 0{,}01 = -\frac{0{,}09}{N}$$
$$-0{,}002 = -\frac{0{,}09}{N}$$
$$0{,}002 = \frac{0{,}09}{N}$$
$$N = \frac{0{,}09}{0{,}002}$$
$$N = 45$$

El tamaño muestral más pequeño es $N = 45$.

7. En regresión logística mostrar que

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

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

La función σ se define como:

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

Sustituimos esta expresión en la ecuación:

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

Un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado porque si el ejemplo está bien clasificado el valor de $e^{y_n w^T x_n}$ tiende a crecer haciendo que el denominador de la fracción crezca, luego el valor del gradiente tenderá a ser un número pequeño. Por el contrario, si el ejemplo está mal clasificado el valor de $e^{y_n w^T x_n}$ tenderá a 0, por lo que el denominador será un número pequeño, lo cual hace que el gradiente tenga un valor más grande que si está bien clasificado.

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

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

Argumentar que el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\nu = 1$.

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 sobrejaustar será mayor o menor? (Ayuda: analizar los detalles que influencian el sobreajuste).

   Si $\mathcal{H}$ es fija e incrementamos la complejidad de f, el ruido determinista en general aumentará dado que habrá puntos de la función f que no podrán ser ajustados por las funciones de la clase $\mathcal{H}$ y ,de igual manera, la tendencia a sobreajustar sería menor. Si f es fija y decrementamos la complejidad de $\mathcal{H}$ estaríamos ante un caso análogo, dado que aumentar la complejidad de f, en general, tiene el mismo efecto que decrementar la complejidad de $\mathcal{H}$; por lo que la tendencia de sobreajuste sería menor y el ruido determinista tendería a subir.

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

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

que define relaciones entre las w_i (La matriz Γ_i se denomina regularizador 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 regularizadores de Tikhonov puede hacerse a través de las propiedades algebraicas de las matrices Γ.

Bonus:

B1. Considerar la matriz hat $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$. Sea una matriz $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ invertible. Mostrar que traza(H) $= d + 1$, donde traza significa la suma de los elementos de la diagonal principal. (+1 punto).

Para demostrarlo partimos de que $(X^T X)$ es una matriz de dimensión $(d+1)(d+1)$ y de la propiedad $traza(AB) = traza(BA)$.

$$traza(H) = traza(X(X^T X)^{-1} X^T) = traza((X^T X)^{-1} X^T X) = traza(I_{d+1}) = d + 1$$