Aprendizaje automático. Cuestionario de teoría 2

Jacinto Carrasco Castillo

15 de mayo de 2016

Cuestión 1. Sean $\mathbf{x}$ e $\mathbf{y}$ dos vectores de observaciones de tamaño N. Sea

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

la covarianza de dichos vectores, donde $\bar{z}$ 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 asociadas 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.

Solución 1. Al estar X dada por columnas, la notaremos:

$$X = (x_{ij})_{\substack{i=1,\dots,N \\ j=1,\dots,d}} = \begin{pmatrix} x_1 & x_2 & \dots & x_d \end{pmatrix}$$

Para definir la matriz de covarianzas nos hará falta también el vector de medias por columnas:

$$\mu = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_d \end{pmatrix}$$

Entonces, la matriz de covarianzas será:

$$\begin{aligned} cov(X) = (cov(x_i, x_j))_{i,j=1,\dots,d} &= \left(\frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)\right)_{i,j=1,\dots,d} = \\ &= \left(\frac{1}{N}\sum_{k=1}^{N}(x_{ki}x_{kj} - \bar{x}_i x_{kj} - \bar{x}_j x_{ki}) + \bar{x}_i\bar{x}_j\right)_{i,j=1,\dots,d} = \\ &= \left(\frac{1}{N}x_i^T x_j\right)_{i,j=1,\dots,d} - (\bar{x}_i\bar{x}_j)_{i,j=1,\dots,d} \end{aligned}$$

Donde hemos usado que $\frac{1}{N}\sum_{k=1}^{N} \bar{x}_i x_{kj} = \bar{x}_i \bar{x}_j$. Ahora observamos que en el primer término de la suma estamos multiplicando escalarmente columnas, lo que significa que estamos multiplicando X^T por X. Entonces nos queda:

$$cov(X) = \frac{1}{N} X^T X - \mu^T \mu$$

Cuestión 2. Considerar 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 natural k

Solución 2.

a Para ver que H es simétrica, la comparamos con su traspuesta:

$$\begin{aligned} H^T &= (X(X^T X)^{-1} X^T)^T = (X^T)^T (X^T X)^{-T} X^T = \\ &= X(X^T X)^{-1} X^T = H \end{aligned}$$

b Lo probamos por inducción. Para $k = 1$ es obviamente cierto. Supuesto cierto para k, lo probamos para $k+1$:

$$\begin{aligned} H^{k+1} = H^k H \overset{hip.ind.}{=} HH = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = \\ = X(X^T X)^{-1} X^T = H \end{aligned}$$

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).

Solución 3. Lo plantearemos como un problema con restricciones y lo resolveremos usando multiplicadores de Lagrange. La función a minimizar será la distancia entre un punto en $\mathbb{R}^2$ y (x_1, y_1), $f(x, y) = \sqrt{(x - x_1)^2 + (y - y_1)^2}$, sujeto a que el punto esté en la recta $ax + by + d = 0$. Por tanto, creamos la función $\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda g(x, y) = f(x, y) + \lambda(ax + by + d)$. Para hallar el punto más cercano a (x_1, y_1) en la recta, hallaremos el punto que cumpla $\bigtriangledown_{(x,y,\lambda)} \mathcal{L}(x, y, \lambda) = 0$

$$\frac{\partial \mathcal{L}}{\partial x}(x, y, \lambda) = \frac{2(x - x_1)}{\sqrt{(x - x_1)^2 + (y - y_1)^2}} + a\lambda = 0 \tag{3.1}$$

$$\frac{\partial \mathcal{L}}{\partial y}(x, y, \lambda) = \frac{2(y - y_1)}{\sqrt{(x - x_1)^2 + (y - y_1)^2}} + b\lambda = 0 \tag{3.2}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda}(x, y, \lambda) = ax + by + d = 0 \tag{3.3}$$

Si igualamos la primera y la segunda ecuación obtenemos

$$\frac{x - x_1}{a} = \frac{y - y_1}{b};\ x = \frac{a}{b}(y - y_1) + x_1 \tag{3.4}$$

Si sustituimos x en (3.3):

$$\frac{a^2}{b}(y - y_1) + ax_1 + by + d = 0$$

$$y = \frac{b(\frac{a^2}{b}y_1 - ax_1 - d)}{a^2 + b^2} = \frac{a^2 y_1 - abx_1 - bd}{a^2 + b^2} \tag{3.5}$$

Por lo que la solución resulta, sustituyendo y en (3.4):

$$x_0 = \frac{b^2 x_1 - aby_1 - ad}{a^2 + b^2}$$

$$y_0 = \frac{a^2 y_1 - abx_1 - bd}{a^2 + b^2}$$

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

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

$$\textit{sujeto a } A\mathbf{z} \leq \mathbf{b}$$

donde $\mathbf{c}$ y $\mathbf{b}$ son vectores y A es una matriz.

1. 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.

2. Formular un problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, identifique quiénes son $A, \mathbf{z}, \mathbf{b}$ y $\mathbf{c}$ para este caso.

Solución 4.

a Que los datos sean linealmente separables significa que $\exists w$ tal que $sign(w^T x_n) = y_n \ \forall n$. Esto es:

- Si $y_n = -1$, $w^T x_n < 0 \Rightarrow y_n w^T x_n > 0$
- Si $y_n = 1$, $w^T x_n > 0 \Rightarrow y_n w^T x_n > 0$

Luego $\exists w$ *tal que* $y_n w^T x_n > 0$

b La expresión del problema de programación lineal asociado a la clasificación se ve más clara si pensamos que los datos no son separables, aunque obviamente llegaremos a la solución si lo fuesen.
Supongamos $\mathbf{x}, \mathbf{w}$ de dimensión d. Para cada dato x_i, suponiendo que no son linealmente separables, podríamos encontrar $\xi_i > 0$ para el que se cumpla $y_i \mathbf{w}^T x_i + \xi_i > 0$, es decir, una holgura que nos deje en la situación del apartado previo. Sea ξ el vector de dimensión N con los ξ_i. Llamamos $\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \xi \end{pmatrix}$. Entonces, hallar $\min_{\mathbf{z}} \sum_{i=0}^{N} \xi_i$, es decir, el mínimo error (si los datos fuesen separables este error sería 0), equivale a hallar $\min_{\mathbf{z}} \underbrace{(0, \overset{d)}{\ldots}, 0, 1, \overset{N)}{\ldots}, 1)^T}_{\mathbf{c}^T} \mathbf{z}$. Ahora resulta sencillo ver qué transformación hay que hacerle a la restricción del sistema de ecuaciones de la programación lineal:

$$\underbrace{\begin{pmatrix} -y_1 x_1^{(1)} & \ldots & -y_1 x_1^{(d)} & -1 & 0 & \overset{N)}{\ldots} & 0 \\ -y_2 x_2^{(1)} & \ldots & -y_2 x_2^{(d)} & 0 & -1 & \overset{N)}{\ldots} & 0 \\ \vdots & \ddots & \vdots & \vdots & & \ddots & \vdots \\ -y_N x_N^{(1)} & \ldots & -y_N x_N^{(d)} & 0 & 0 & \overset{N)}{\ldots} & -1 \\ 0 & \ldots & 0 & -1 & 0 & \overset{N)}{\ldots} & 0 \\ \vdots & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & 0 & \overset{N)}{\ldots} & -1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \mathbf{w}^{(1)} \\ \mathbf{w}^{(2)} \\ \vdots \\ \mathbf{w}^{(d)} \\ \xi_1 \\ \vdots \\ \xi_N \end{pmatrix}}_{\mathbf{z}} \leq \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{b}}$$

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

Solución 5. Para este ejercicio consideraremos que la función con ruido que queremos ajustar es $f(x) + \varepsilon$. Entonces, calcular el error esperado fuera de la muestra será:

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

Si, al igual que para el caso sin ruido, introducimos la esperanza en (D) sumando y restando y desarrollamos el cuadrado de dentro de la esperanza en x obtenemos:

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

Bien, este no es el resultado esperado, sin embargo hemos pasado algo por alto, el error fuera de la muestra depende de la distribución que siga el error introducido en la función objetivo, con lo que en realidad el error $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})}]$ no es $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(g^{(\mathcal{D})} - f(x) - \varepsilon)^2]]$ sino $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{x,\varepsilon}(g^{(\mathcal{D})} - f(x) - \varepsilon)^2]]$. Entonces, si seguimos por la cuenta anterior haciendo también la esperanza con respecto de ε, partiendo de la hipótesis de que ε tiene media 0 y varianza σ^2 obtenemos, usando que ε es independiente de x y $f(x), \bar{g}(x)$ lo son de ε:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{x,\varepsilon}(g^{(\mathcal{D})} - f(x) - \varepsilon)^2]] = \\
&= var + bias + \mathbb{E}_{\varepsilon}[\varepsilon^2] + \mathbb{E}_{\varepsilon}[\varepsilon]\mathbb{E}_x[\bar{g}(x) - f(x)] = \\
&= var + bias + \sigma^2
\end{aligned}$$

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?

Solución 6. Por el ejercicio 2 del apartado de regresión tenemos que el error esperado de regresión lineal respecto a $\mathcal{D}$ es:

$$\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$$

Entonces, sustituimos con los datos del enunciado imponiendo $E_{in} >$ 0,008:

$$0{,}008 < \sigma^2 \left(1 - \frac{d+1}{N}\right) = 0{,}01 \left(1 - \frac{9}{N}\right);$$

$$0{,}8 < 1 - \frac{9}{N}; \quad \frac{9}{N} < 0{,}2; \quad 45 < N$$

Con lo que llegamos a que el tamaño muestral debe ser mayor de 45.

Cuestión 7. En regresión logística mostrar que

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

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

Solución 7.

$$\bigtriangledown_{\mathbf{w}} E_{in}(\mathbf{w}) = \left(\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}_i} \log(1 + e^{-y_n \mathbf{w}^T x_n})\right)_{i=1,\dots d} =$$

$$= \left(\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}_i} \log(1 + e^{-y_n \mathbf{w}^T x_n})\right)_{i=1,\dots d} =$$

$$= \frac{1}{N}\left(\sum_{n=1}^{N} \frac{-y_n x_{ni} e^{-y_n \mathbf{w}^T x_n}}{1 + e^{-y_n \mathbf{w}^T x_n}}\right)_{i=1,\dots d} = \frac{1}{N}\left(\sum_{n=1}^{N} \frac{-y_n x_{ni}}{1 + e^{y_n \mathbf{w}^T x_n}}\right)_{i=1,\dots d} =$$

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

Para argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado, vemos cómo $1 + e^{y_n \mathbf{w}^T x_n} > 1 + e^{y_m \mathbf{w}^T x_m}$ si el dato m está clasificado y el dato n no está clasificado, puesto que $y_n \mathbf{w}^T x_n < 0$ y entonces $e^{y_n \mathbf{w}^T x_n} < e^{y_m \mathbf{w}^T x_m}$. Esto hace que el numerador sea más grande en el segundo caso y por tanto el gradiente y la contribución es menor si el dato está bien clasificado. No se puede hacer un argumento más elaborado ya que es posible que, aún estando el dato m bien etiquetado, su contribución al gradiente sea mayor en término absoluto que la de un dato mal etiquetado y que, por tener el valor de la norma del punto (que influye en el numerador) sea muy pequeño y por tanto la contribución termine siendo menor pese a no estar clasificado.

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

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

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

Solución 8. Si $-y_n \mathbf{w}^T \mathbf{x}_n < 0$ significa que está bien clasificado, lo que significa que sólo estaremos usando para actualizar los datos mal clasificados. Si le hacemos el gradiente a la función $\mathbf{e}_n$, obtenemos:

$$\bigtriangledown_{\mathbf{w}} \mathbf{e}_n = \begin{cases} -y_n \mathbf{x}_n & si\ y_n \mathbf{w}^T \mathbf{x}_n < 0 \\ 0 & si\ y_n \mathbf{w}^T \mathbf{x}_n > 0 \end{cases}$$

Y entonces la regla de actualización $\mathbf{w}_{old} = \mathbf{w} - \eta \bigtriangledown_{\mathbf{w}} \mathbf{e}_n$, la regla de actualización del $\mathbf{w}$ en `SGD` es justo la regla de actualización del algoritmo `PLA`.

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

1. Suponer que $\mathcal{H}$ es fija y que incrementamos la complejidad de f.
2. 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 influyen en el sobreajuste)

Solución 9.

1. Si dejamos $\mathcal{H}$ fija e incrementamos la complejidad de f el ruido determinista aumentará al salir f del espacio de funciones de $\mathcal{H}$, habiendo una mayor diferencia cada vez entre la función en $\mathcal{H}$ que mejor aproxima y f. La tendencia a sobreajustar aumentará, ya que, aunque la varianza se mantenga constante (no hemos modificado $\mathcal{H}$, luego la diferencia entre las posibles $g^{(\mathcal{D})}$ y $\bar{g}$ será la misma), pero sí aumentarán las diferencias entre $\bar{g}$ y f.

2. Si mantenemos la complejidad de f y decrementamos la de $\mathcal{H}$ aumentará el ruido determinista, ya que nos iremos alejando de la posibilidad de que $\mathcal{H}$ contenga a f. En cambio, la posibilidad de sobreajuste sí disminuye, ya que aunque es cierto como en el caso anterior que el sesgo aumenta, ahora sí disminuye la varianza, puesto que al haber menos funciones g donde escoger, la diferencia entre la función media y la función dada por los datos será menor.

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 se denomina regularizador de Tikhonov)

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

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

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

Solución 10.

a Para este primer caso queremos que la norma de $\mathbf{w}$ sea menor o igual que C. Esto es $\mathbf{w}^T\mathbf{w} \leq C$, lo que significa que $\Gamma^T\Gamma = I$, es decir, $\Gamma^T = \Gamma^{-1}$, con lo que necesitamos que Γ sea ortogonal.

b Si es la norma de la suma al cuadrado lo que tenemos que acotar, podemos tomar $\Gamma = (1\overset{(Q_f)}{\ldots}1)$ y obtenemos para $\mathbf{w}^T\Gamma^T$ y $\Gamma\mathbf{w}$ la la norma de la suma, con lo que la condición nos queda como esperábamos.

Como se ve en el primer apartado, las propiedades algebraicas de Γ influyen decisivamente en los regularizadores de Tickhonov.

Cuestión Opcional 1. 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.

Solución Opcional 1. Este resultado es inmedito usando la propiedad de la traza:

$$traza(AB) = traza(BA)$$

aplicándola a H:

$$traza(H) = traza(X(X^TX)^{-1}X^T) = traza(X^TX(X^TX)^{-1}) = traza(I)$$

Ahora, $X^TX \in \mathcal{M}_{(d+1),(d+1)}$, luego la matriz identidad obtenida pertenece a $\mathcal{M}_{(d+1),(d+1)}$ y por tanto su traza es $d+1$.