Aprendizaje automático
Cuestionario 2

Alejandro García Montoro
agarciamontoro@correo.ugr.es

15 de mayo de 2016

1. Ejercicios

Ejercicio 1. *Sean x e y dos vectores de observaciones de tamaño N. Sea*

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

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 $\text{cov}(X)$ en función de la matriz X.

Solución. ss

Ejercicio 2. *Considerar 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.*

1. *Mostrar que H es simétrica*
2. *Mostrar que $H^K = H$ para cualquier entero K.*

Solución. Decir que H es simétrica es equivalente a decir que es igual a su traspuesta. Por tanto, nos basta comprobar que $H^T = H$. Comprobémoslo:

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

donde hemos usado las siguientes propiedades de la traspuesta: $(AB)^T = B^T A^T$, $(A^T)^{-1} = (A^{-1})^T$ y $(A^T)^T = A$.

Para ver que $H^K = H$ para cualquier entero K, vamos primero a estudiar el cuadrado de H:

$$\begin{aligned} H^2 &= \left(X(X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}X^T\right) = \\ &= X(X^TX)^{-1}\left((X^TX)(X^TX)^{-1}\right)X^T = \\ &= X(X^TX)^{-1}I_{d+1}X^T = X(X^TX)^{-1}X^T = \\ &= H \end{aligned}$$

donde hemos denotado como I_{d+1} a la matriz identidad de dimensión $d+1$.

Tenemos así que $H^2 = H$. Ahora basta aplicar inducción para terminar el razonamiento:

- Para el caso base $K = 2$ está probado: $H^2 = H$.
- Lo suponemos cierto para $K-1$; esto es, $H^{K-1} = H$.
- Lo vemos para K, donde usamos la hipótesis de inducción primero y el caso base después:

$$H^K = H^{K-1}H = HH = H^2 = H$$

Concluimos así que $H^K = H$ para todo K.

Ejercicio 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. Estamos ante un problema de minimización con restricciones, así que vamos a usar la técnica de los multiplicadores de Lagrange para resolverlo.

La función a minimizar es la siguiente:

$$g(x, y) = d((x, y), (x_1, y_1)) = \sqrt{(x - x_1)^2 + (y - y_1)^2}$$

Como lo que nos interesa es el punto donde se alcanza el mínimo y no el valor de la función en ese punto, podemos considerar como función a minimizar el cuadrado de g, lo que nos facilitará los cálculos más adelante.

La restricción es la siguiente:

$$ax + by + d = 0$$

donde suponemos a,b y d conocidos y tales que $a \neq 0$ y $b \neq 0$.

El lagrangiano de este problema es el siguiente:

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

cuyas derivadas parciales son:

$$\frac{\partial}{\partial x}\mathcal{L}(x,y,\lambda) = 2(x-x_1) - \lambda a$$
$$\frac{\partial}{\partial y}\mathcal{L}(x,y,\lambda) = 2(y-y_1) - \lambda b$$
$$\frac{\partial}{\partial \lambda}\mathcal{L}(x,y,\lambda) = -(ax+by+d)$$

Igualando las tres derivadas parciales a cero obtenemos el sistema de tres ecuaciones con tres incógnitas —a saber, x, y y λ— que tenemos que resolver:

$$2(x-x_1) = \lambda a$$
$$2(y-y_1) = \lambda b$$
$$ax+by+d = 0$$

Despejando lambda de la primera ecuación, de donde obtenemos que $\lambda = \frac{2(x-x_1)}{a}$, y sustituyendo en la segunda, tenemos lo siguiente:

$$y - y_1 = \frac{b}{a}(x-x_1) \qquad \text{(Usamos que } a \neq 0\text{)}$$
$$ax+by+d = 0$$

Despejamos, por ejemplo, y de la primera ecuación, de donde obtenemos $y = y_1 + \frac{b}{a}(x-x_1)$, y sustituimos en la segunda, de donde podemos obtener ya el valor de x:

$$ax + b(y_1 + \frac{b}{a}(x-x_1)) + d = 0$$
$$ax + by_1 + \frac{b^2}{a}x - \frac{b^2}{a}x_1 + d = 0$$
$$x = \frac{-by_1 + \frac{b^2}{a}x_1 - d}{a + \frac{b^2}{a}} \qquad \text{(Usamos que } b \neq 0\text{)}$$

Sustituyendo por último en la ecuación que teníamos para y, obtenemos que el punto (x_0, y_0) sobre la línea $ax+by+d=0$ que está más cerca del (x_1, y_1) es:

$$(x_0,y_0) = (\frac{-by_1 + \frac{b^2}{a}x_1 - d}{a+\frac{b^2}{a}}, y_1 + \frac{b}{a}(\frac{-by_1 + \frac{b^2}{a}x_1 - d}{a+\frac{b^2}{a}} - x_1))$$

Ejercicio 4.

Consideremos el problema de optimización lineal con restricciones definido por

$$\min_z\{c^T z\} \text{ sujeto a } Az \leq b$$

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

1. *Para un conjunto de datos linealmente separable mostrar que para algún w se debe de verificar la condición $y_n w^T x_n > 0$ para todo (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, $\boldsymbol{z}$, $\boldsymbol{b}$ y $\boldsymbol{c}$ para este caso.*

Solución. Razonemos por reducción al absurdo: supongamos que para todo w existe al menos algún (x_n, y_n) tal que $y_n w^t x_n \leq 0$. Esta es la definición de conjunto no linealmente separable, ya que para cualquier hiperplano hay al menos una muestra mal clasificada. Esto entra en contradicción con la hipótesis de que el conjunto es linealmente separable, luego concluimos que existe un w tal que $y_n w^t x_n < 0 \ \ \forall (x_n, y_n)$.

Para formular el problema de optimización lineal, tenemos que tener en cuenta que lo que queremos minimizar —en el caso de un conjunto linealmente separable queremos que sea nulo— es el error dentro de la muestra, $E_{in}(w)$

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

Solución. En clase se vio esta fórmula para funciones con ruido. Vamos aquí reproducir aquel desarrollo usando esta vez una función con general con ruido

$$y(x) = f(x) + \varepsilon$$

donde ε es una variable aleatoria de media cero y varianza σ^2 que modela el ruido. Veamos el desarrollo:

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

donde en la penúltima igualdad hemos usado que la media de la variable que modela el ruido es cero; es decir, que $\mathbb{E}_x[\varepsilon] = 0$ y, en la última, la definición de varianza; esto es: $\mathbb{E}_x[\varepsilon^2] = \mathbb{E}_x[(\varepsilon - 0)^2] = \mathbb{E}_x[(\varepsilon - \mathbb{E}_x[\varepsilon])^2] = \mathrm{var}(\varepsilon) = \sigma^2$.

Ejercicio 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. Sabemos, por el ejercicio citado, que el valor esperado de E_{in} es:

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

Si queremos que este valor sea mayor que $\varepsilon = 0{,}008$, basta imponerlo y despejar N para ver qué tamaño muestral necesitamos:

$$\begin{aligned}
\sigma^2\left(1-\frac{d+1}{N}\right) &> \varepsilon \\
1-\frac{d+1}{N} &> \frac{\varepsilon}{\sigma^2} \\
1-\frac{\varepsilon}{\sigma^2} &> \frac{d+1}{N} \\
N\left(1-\frac{\varepsilon}{\sigma^2}\right) &> d+1 \\
N &> \frac{d+1}{1-\frac{\varepsilon}{\sigma^2}}
\end{aligned}$$

Concluimos que el mínimo tamaño muestral para que el valor esperado de E_{in} sea mayor que ε es de $N = \lceil \frac{d+1}{1-\frac{\varepsilon}{\sigma^2}} \rceil$. Sustituyendo los valores indicados, obtenemos que este número es:

$$N = \lceil \frac{8+1}{1-\frac{0{,}008}{0{,}1^2}} \rceil = 45$$

Ejercicio 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.

Solución. En regresión logística, el error está definido como sigue:

$$E_{in}(w) = \frac{1}{N}\sum_{n=0}^{N}\ln(1+e^{-y_n w^T x_n})$$

Por tanto, su gradiente es:

$$\begin{aligned}
\nabla E_{in}(w) &= \frac{\partial}{\partial w}\left(\frac{1}{N}\sum_{n=0}^{N}\ln(1+e^{-y_n w^T x_n})\right) = \\
&= \frac{1}{N}\sum_{n=0}^{N}\frac{\partial}{\partial w}\left(\ln(1+e^{-y_n w^T x_n})\right) = \\
&= \frac{1}{N}\sum_{n=0}^{N}\frac{-y_n x_n e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} = \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Multiplico por } \tfrac{e^{y_n w^T x_n}}{e^{y_n w^T x_n}}\text{)} \\
&= \frac{1}{N}\sum_{n=0}^{N}\frac{-y_n x_n}{e^{y_n w^T x_n}+1} = \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Func. logística: } \sigma(t) = \tfrac{1}{1+e^{-t}}\text{)} \\
&= \frac{1}{N}\sum_{n=0}^{N} -y_n x_n \sigma(-y_n w^T x_n)
\end{aligned}$$

Fijándonos en la penúltima igualdad,

$$\nabla E_{in}(w) = \frac{1}{N}\sum_{n=0}^{N}\frac{-y_n x_n}{e^{y_n w^T x_n}+1}$$

es claro que son los ejemplos mal clasificados los que tienen un gran peso en el error.

Que un ejemplo esté bien clasificado implica que el exponente $y_n w^T x_n$ es positivo, ya que y_n y $w^T x_n$ tienen el mismo signo. Por tanto, la exponencial es siempre mayor que uno y, a su vez, el denominador será siempre mayor que 2.

Que un ejemplo esté mal clasificado implica que el exponente $y_n w^T x_n$ es negativo, ya que y_n y $w^T x_n$ tienen signos contrarios. Como la exponencial de un número negativo es siempre menor que uno, el denominador será siempre menor que 2.

Hemos visto así que un ejemplo mal clasificado tiene siempre un denominador más pequeño que el de un ejemplo bien clasificado; es decir, el sumando asociado a un ejemplo mal clasificado es mayor que el de un ejemplo bien clasificado, contribuyendo así en mayor medida al gradiente.

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

$$e_n(w) = \text{máx}(0, -y_n w^T x_n)$$

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

Solución. Tras inicializar los pesos w que definen el hiperplano, el algoritmo PLA recorre todas las muestras (x_n, y_n) y, por cada una de las mal clasificadas, ejecuta la siguiente regla de actualización:

$$w \leftarrow w + y_n x_n$$

El algoritmo SGD, por otro lado, ejecuta la siguiente regla de actualización *para todas* las muestras, estén o no mal clasificadas:

$$w \leftarrow w - \eta \nabla e_n(w)$$

y donde e_n es el error definido en un punto.

Ahora bien, si tomamos como error el indicado en el enunciado, su gradiente es el siguiente:

$$\begin{aligned}\nabla e_n(w) &= \frac{\partial}{\partial w} \text{máx}(0, -y_n w^T x_n) = \text{máx}(\frac{\partial}{\partial w}(0), \frac{\partial}{\partial w}(-y_n w^T x_n)) = \\ &= \text{máx}(0, -y_n x_n)\end{aligned}$$

Estudiemos ahora cómo se comporta ese gradiente. Sea (x_n, y_n) una muestra mal etiquetada —es decir, el signo de y_n y de x_n es distinto—, entonces:

$$\begin{aligned}\text{sign}(y_n) \neq \text{sign}(x_n) \Rightarrow y_n x_n < 0 \Rightarrow -y_n x_n > 0 \Rightarrow \\ \Rightarrow \nabla e_n(w) = -y_n x_n\end{aligned}$$

Sea ahora (x_n, y_n) una muestra bien etiquetada —es decir, el signo de y_n es igual al signo de x_n—, luego:

$$\begin{aligned}\text{sign}(y_n) = \text{sign}(x_n) \Rightarrow y_n x_n > 0 \Rightarrow -y_n x_n < 0 \Rightarrow \\ \Rightarrow \nabla e_n(w) = 0\end{aligned}$$

Concluimos entonces que la regla de actualización en SGD, con $\eta = 1$, es la siguiente para todas las muestras:

$$w \leftarrow w - \begin{cases} 0 & \text{si la muestra } x_n y_n \text{ está bien etiquetada} \\ -y_n x_n & \text{si la muestra } x_n y_n \text{ está mal etiquetada} \end{cases}$$

Por tanto, la actualización real se hace si y sólo si la muestra está mal etiquetada, y su expresión final es:

$$w \leftarrow w + y_n x_n$$

es decir, el SGD con $\eta = 1$ y error $e_n(w) = \text{máx}(0, -y_n w^T x_n)$ tiene exactamente el mismo comportamiento que el algoritmo PLA.

Ejercicio 9. *El ruido determinista depende de $\mathcal{H}$, ya que 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 influencian el sobreajuste.

Solución. El ruido determinista, por definición, es la consecuencia de ajustar funciones objetivo f con funciones de un modelo $\mathcal{H}$ más simple que f. La diferencia entre lo que $\mathcal{H}$ intenta ajustar y los datos reales de f tiene al final el mismo comportamiento en el resultado que el ruido estocástico. El primero, sin embargo, no depende de cada toma de datos sino de la función que los genera, luego es constante entre conjuntos de datos extraídos de una misma función.

Teniendo en cuenta la definición del ruido determinista, es claro que en el primer escenario este subirá. Al tener un modelo fijo de funciones que intentan aproximar una función f cada vez más compleja, la complejidad que el modelo es incapaz de ajustar terminará actuando como ruido. Por tanto, el sobreajuste será mayor: el aumento de ruido, ya sea determinista o estocástico —el modelo no sabe de qué tipo es el ruido, ni siquiera si lo hay—, siempre influye en el sobreajuste.

El segundo escenario es algo diferente. Si bien la primera pregunta tiene la misma respuesta, ya que la complejidad de la función objetivo será mayor que la de la clase de funciones que intenta ajustarla y, por tanto el ruido determinista va a aumentar, en el caso del sobreajuste el comportamiento no es tan claro.

Sabemos que el sobreajuste ocurre cuando un modelo intenta aproximar una función objetivo de tal manera que llega a ajustar el ruido —ya sea este estocástico o determinista—. Sin embargo, para que esto ocurra, el modelo debe tener los suficientes grados de libertad como para aprender del ruido. Si el modelo es más simple, intentará siempre ajustarse al comportamiento general de la muestra y, aunque el error dentro de ella sea mayor que con modelos más complejos, la falta de libertad puede influir positivamente en el error fuera de la muestra: al haber captado el comportamiento general y no haber podido aprender del ruido —por falta de libertad, no porque este no exista—, el error fuera de la muestra puede ser mejor que con modelos más complejos, cuyos grados de libertad terminarán influyendo negativamente.

Por tanto, podemos concluir que al intentar ajustar una misma función objetivo con modelos cada vez más simples, la tendencia a sobreajustar será menor.

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

$$w^T \Gamma^T \Gamma w \leq C$$

que define relaciones entre las w_i —la matriz Γ_i se denomina regularizador de Tikhonov—.

1. *Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$*

2. *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.

2. Bonus

Bonus 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. (+1 punto)*

Solución. Sabemos que traza(AB) = traza(BA) con A y B matrices cualesquiera. Por tanto:

$$\begin{aligned}
\text{traza}(H) &= \text{traza}(\underbrace{X}_{\text{A}}\underbrace{(X^TX)^{-1}X^T}_{\text{B}}) = \\
&= \text{traza}(\underbrace{(X^TX)^{-1}X^T}_{\text{B}}\underbrace{X}_{\text{A}}) = \\
&= \text{traza}\left(\left((X^TX)^{-1}\right)\left(X^TX\right)\right) = \\
&= traza(I_{d+1}) = d+1
\end{aligned}$$