

## Cuestionario 2

Luis Suárez Lloréns

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

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

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 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  $X$ .

**Respuesta.** Siendo  $X$  la matriz, supongamos que  $\bar{X}$  una matriz que tiene como valor de las columnas la media de las columnas de  $X$ . Entonces:

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

Pues con esta operación, en la posición  $(i, j)$  de  $\text{cov}(X)$  obtenemos:

$$\frac{1}{N} \sum_{k=1}^N (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)$$

siendo  $\bar{x}_i$  la media de la columna  $i$ .

Falta definir entonces  $\bar{X}$ :

$$\bar{X} = \frac{1}{N} 1_{n \times n} X$$

Siendo  $1_{n \times n}$  la matriz con todos sus valores 1, y  $n$  es el número de datos que tenemos por variable.

**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 entero  $K$ .

**Respuesta. a)**

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

b) Si  $H^2 = H$ , entonces por inducción, tendríamos que  $H^K = H$ . Por tanto, sólo vamos a demostrar  $H^2 = H$ .

$$H^2 = HH = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = \\ X(X^T X)^{-1} (X^T X (X^T X)^{-1}) X^T = X(X^T X)^{-1} X^T = H$$

**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  $(x_1, y_1)$ .

**Respuesta.**

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

$$\text{Min}_{\mathbf{z}} \mathbf{c}^T \mathbf{z} \\ \text{Sujeto a } A\mathbf{z} \leq \mathbf{b}$$

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

- 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.
- 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.

**Respuesta.**

**Cuestión 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).

**Respuesta.**

**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?

**Respuesta.**

**Cuestión 7.** En regresión logística mostrar que

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

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

**Respuesta.** Primero, destacar que:

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

Luego de las dos expresiones que tenemos, aplicando la segunda expresión de  $\sigma(x)$  a la segunda de  $\nabla E_{in}$  obtenemos directamente la primera expresión de

$\nabla E_{in}$ .

Por tanto, nos queda calcular directamente la expresión de  $\nabla E_{in}$  y ver que es igual que cualquiera de las dos:

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

$$\nabla E_{in} = \frac{1}{N} \sum_{n=1}^N \frac{1}{1 + e^{-y_n w^T x_n}} \left(1 + e^{-y_n w^T x_n}\right)' =$$

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

Sustituimos por la primera expresión de  $\sigma(-y_n w^T x_n)$  y obtenemos:

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

Por último, si un dato está bien clasificado,  $y_n w^T x_n$  es positivo, y si lo clasifica mal, el número es negativo. Si observamos la primera expresión de  $\nabla E_{in}$ , vemos que, si  $y_n w^T x_n$  es positivo, dividimos  $y_n x_n$  por un número mayor que cuando es negativo.

Por tanto, si un dato está mal clasificado, tendrá más importancia —pues dividimos por un número más pequeño que cuando está bien clasificado—.

**Cuestión 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  $\nu = 1$ .

**Respuesta.** Para empezar, el algoritmo PLA actualiza sus pesos en función de un único dato. Por tanto, si fuera un SGD, el tamaño de "mini-batch" sería de 1.

La función de actualización del PLA es la siguiente:

- Si está bien clasificado, no se actualiza.
- Si está mal clasificado, se actualiza usando  $y_n w^T x_n$ .

En el caso del SGD presentado en el ejercicio, si un dato está bien clasificado,  $-y_n w^T x_n$  es negativo y por tanto, el máximo es 0. Es decir, no se actualiza.

En el caso de un dato mal clasificado,  $-y_n w^T x_n$  es positivo. Sólo que da por ver la diferencia de signo, pero en el caso del SGD se resta el valor dado por  $\mathbf{e}_n$ , luego el valor es el mismo,  $y_n w^T x_n$ .

Por último, al ser el mismo valor exactamente,  $\nu$  tiene que ser 1.

**Cuestión 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 influyen el sobreajuste).

**Respuesta. a)** Al aumentar la complejidad de la función objetivo, aumentamos el ruido determinista. Esto se debe a que la mejor aproximación de la clase  $\mathcal{H}$  no puede modelar a la función objetivo  $f$ .

Como hemos dicho, se aumenta el ruido, dando como sabemos un mayor sobreajuste, pues la función obtenida intentará aprender los valores obtenidos, que tienen errores, y no la mejor función posible de  $\mathcal{H}$ .

**b)** También se aumenta el ruido determinista, pues al reducirse la complejidad de  $\mathcal{H}$  disminuye, y por tanto la mejor función de  $\mathcal{H}$  representa cada vez peor la función real.

Ahora bien, al tener un modelo más sencillo, la función que obtenemos de  $\mathcal{H}$  sólo obtiene las características generales de la función real. Esto hace que, pese a tener más error determinista, la función obtenida tiene más capacidad de generalización.

Visto desde otra perspectiva, ir hacia un modelo más sencillo, aumenta un poco el bias, pero disminuye mucho la varianza, consiguiendo un menor error y un menor sobreajuste.

**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  $\Gamma_i$  se denomina regularizados de Tikhonov)

- a) Calcular  $\Gamma$  cuando  $\sum_{q=0}^Q w_q^2 \leq C$
- b) Calcular  $\Gamma$  cuando  $(\sum_{q=0}^Q w_q)^q \leq C$

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

**Respuesta.** Siendo  $w$ , un vector columna de  $n$  elementos:

- **a)** La matriz que tenemos que seleccionar es la identidad  $I_{n \times n}$ .
- **b)** La matriz que tenemos que seleccionar es la matriz cuadrada  $n \times n$  que tiene la primera fila rellena de 1 y el resto 0.

En cuanto al estudio de los regularizadores de Tikhonov, es claro que el único factor que puede cambiar son las matrices  $\Gamma$ . Por tanto, su estudio y propiedades depende directamente de las propiedades de la matriz que seleccionamos como  $\Gamma$ .

## Bonus

**Bonus.** 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  $\text{traza}(\hat{H}) = d + 1$ , donde traza significa la suma de los elementos de la diagonal principal.

**Respuesta.**