Aprendizaje automático: Cuestionario 2.

Cristina Zuheros Montes.

15/05/2016.

Todas las preguntas tienen el mismo valor

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

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

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 cov(X) en función de la matriz X

Solución
Vamos a considerar la matriz X definida como una matriz cuyas columnas representa vectores de observaciones:

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{21} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{n2} \\ ... & ... & ... & ... \\ x_{1N} & x_{2N} & ... & x_{nN} \end{pmatrix}$$

donde $(\mathbf{xi})^T = (x_{i1}, x_{i2}, ..., x_{iN})$ son los vectores de observaciones.

Como la matriz de covarianza de X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas, podemos definir la covarianza de dicha matriz como:

$$\mathrm{cov}(\mathrm{X}) = \begin{pmatrix} \mathrm{cov}(\mathbf{x1}, \mathbf{x1}) & \mathrm{cov}(\mathbf{x2}, \mathbf{x1}) & ... & \mathrm{cov}(\mathbf{xn}, \mathbf{x1}) \\ \mathrm{cov}(\mathbf{x1}, \mathbf{x2}) & \mathrm{cov}(\mathbf{x2}, \mathbf{x2}) & ... & \mathrm{cov}(\mathbf{xn}, \mathbf{x2}) \\ ... & ... & ... & ... \\ \mathrm{cov}(\mathbf{x1}, \mathbf{xn}) & \mathrm{cov}(\mathbf{x2}, \mathbf{xn}) & ... & \mathrm{cov}(\mathbf{xn}, \mathbf{xn}) \end{pmatrix}$$

Es claro que $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \mathrm{cov}(\mathbf{y}, \mathbf{x})$, luego la matriz de covarianzas asociada a la matriz X será simétrica y se puede ser como:

$$\mathrm{cov}(\mathrm{X}) = \begin{pmatrix} \mathrm{cov}(\mathbf{x1}, \mathbf{x1}) & \mathrm{cov}(\mathbf{x1}, \mathbf{x2}) & ... & \mathrm{cov}(\mathbf{x1}, \mathbf{xn}) \\ \mathrm{cov}(\mathbf{x1}, \mathbf{x2}) & \mathrm{cov}(\mathbf{x2}, \mathbf{x2}) & ... & \mathrm{cov}(\mathbf{x2}, \mathbf{xn}) \\ ... & ... & ... & ... \\ \mathrm{cov}(\mathbf{x1}, \mathbf{xn}) & \mathrm{cov}(\mathbf{x2}, \mathbf{xn}) & ... & \mathrm{cov}(\mathbf{xn}, \mathbf{xn}) \end{pmatrix}$$

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

   (a) Mostrar que H es simétrica
   Solución
   Sabemos que una matriz $H \in \mathcal{H}$ es simétrica si verifica $H^T = H$.
   Tenemos $H^T = (X(X^TX)^{-1}X^T)^T = (X^T)^T((X^TX)^{-1})^T(X^T) = X((X^TX)^{-1})^T(X^T)$.
   Como X^TX es invertible, tendremos que la inversa y la traspuesta conmutan, luego podemos afirmar que $H^T = X((X^TX)^T)^{-1}(X^T) = X(X^T(X^T)^T)^{-1}(X^T) = X(X^TX)^{-1}(X^T) = H$ demostrando que, efectivamente, H es simétrica.

(b) Mostrar que $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero K

Solución

Veamos por inducción:

Para k=1, esto es trivial.
Para k=2, veamos que $H^2 = H$:
$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$.
X es una matriz $N \times (d+1)$, luego X^T es una matriz $(d+1) \times N$. Tenemos que $(X^TX)^{-1}(X^TX) = I$ donde I representa a la matriz identidad de dimensión $d+1$. Quedando la igualdad:
$H^2 = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

Supongamos que es cierto para k, veamos que también se cumple para k+1. Es decir, tenemos que $H^k = H$, veamos que $H^{k+1} = H$. También usaremos que para $k = 2$ hemos visto que es cierto.
$H^{k+1} = HH^k = HH = H^2 = H$.

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).

Solución

Partimos de la recta $r := ax + by + d = 0$ cuya pendiente viene dada por $m_r = \frac{-a}{b}$. Buscamos una recta s perpendicular a la recta r y que pase por el punto (x_1, y_1). Por ser perperdicular a r, tendremos que la pendiente de s tiene que ser $m_s = \frac{b}{a}$. La recta s pasará por el punto (x_1, y_1) y tendrá pendiente $m_s = \frac{b}{a}$. Esto nos lleva a que la recta s verificará:

$$a(y - y_1) = b(x - x_1) \rightarrow bx - ay - bx_1 + ay_1 = 0$$

luego la recta s tendrá la forma: $s := bx - ay + (ay_1 - bx_1) = 0$.
Para obtener el punto (x_0, y_0) basta obtener la intersección entre ambas:

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}^{-1} \begin{pmatrix} -d \\ bx_1 - ay_1 \end{pmatrix} = \frac{1}{a^2 + b^2} \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \begin{pmatrix} -d \\ bx_1 - ay_1 \end{pmatrix} \rightarrow$$

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

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

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

$$\min_{\mathbf{z}} \mathbf{c}^{\mathrm{T}}\mathbf{z}$$
$$\text{Sujeto a } \mathrm{A}\mathbf{z} \leq \mathbf{b}$$

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

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

Solución

(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.

Solución

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

Solución
Al estar trabajando con ruido, vamos a definir $y(x) = f(x)+\epsilon$. De modo que $E_{out}(g^D) = \mathbb{E}_{xy}[(g^D(x)-y(x))^2]$. Vamos a probar la igualdad haciendo los siguientes desarrollos (Destacar que usaremos $\overline{g}(x) = \mathbb{E}_D[g^D(x)]$ y propiedades básicas):

$$\mathbb{E}_D[E_{out}(g^D)] =$$
$$\mathbb{E}_D[\mathbb{E}_{xy}[(g^D(x) - y(x))^2]] =$$
$$\mathbb{E}_D[\mathbb{E}_{xy}[g^D(x)^2] - 2\mathbb{E}_{xy}[g^D(x)]\mathbb{E}_{xy}[y(x)] + \mathbb{E}_{xy}[y(x)^2]] =$$
$$\mathbb{E}_{xy}[\mathbb{E}_D[g^D(x)^2]] - 2\mathbb{E}_{xy}[\mathbb{E}_D[g^D(x)]\mathbb{E}_D[y(x)]] + \mathbb{E}_{xy}[\mathbb{E}_D[y(x)^2]] =$$
$$\mathbb{E}_{xy}[\mathbb{E}_D[g^D(x)^2] - 2\overline{g}(x)\mathbb{E}_D[y(x)] + \mathbb{E}_D[y(x)^2]] =$$
$$\mathbb{E}_{xy}[(\mathbb{E}_D[g^D(x)^2] - \overline{g}(x)^2) + (\overline{g}(x)^2 - 2\overline{g}(x)\mathbb{E}_D[y(x)] + \mathbb{E}_D[y(x)^2])] =$$

Para evitar arrastrar tantos valores, vamos a reducir de forma separada la primera y segunda componente de la suma principal. Veamos la primera componente:

$$\mathbb{E}_D[g^D(x)^2] - \overline{g}(x)^2 =$$
$$\mathbb{E}_D[g^D(x)^2] - 2\overline{g}(x)^2 + \overline{g}(x)^2 =$$
$$\mathbb{E}_D[g^D(x)^2 - 2g^D(x)\overline{g}(x) + \overline{g}(x)^2] =$$
$$\mathbb{E}_D[(g^D(x)^2 - \overline{g}(x))^2]$$

Veamos ahora la segunda componente de la suma:

$$\overline{g}(x)^2 - 2\overline{g}(x)\mathbb{E}_D[y(x)] + \mathbb{E}_D[y(x)^2] =$$
$$\overline{g}(x)^2 - 2\overline{g}(x)\mathbb{E}_D[f(x) + \epsilon] + \mathbb{E}_D[(f(x) + \epsilon)^2] =$$
$$\overline{g}(x)^2 - 2\overline{g}(x)\mathbb{E}_D[f(x)] - 2\overline{g}(x)\mathbb{E}_D[\epsilon] + \mathbb{E}_D[f(x)^2] + 2\mathbb{E}_D[f(x)\epsilon] + \mathbb{E}_D[\epsilon^2] =$$
$$(\overline{g}(x)^2 - 2\overline{g}(x)f(x) + f(x)^2) - 2\overline{g}(x)\mathbb{E}_D[\epsilon] + 2\mathbb{E}_D[f(x)\epsilon] + \mathbb{E}_D[\epsilon^2] =$$
$$(\overline{g}(x) - f(x))^2 - 2\overline{g}(x)\mathbb{E}_D[\epsilon] + 2\mathbb{E}_D[f(x)\epsilon] + \mathbb{E}_D[\epsilon^2]$$

Retomando por donde nos habíamos quedado, tenemos:

$$\mathbb{E}_D[E_{out}(g^D)] =$$
$$\mathbb{E}_{xy}[(\mathbb{E}_D[g^D(x)^2] - \overline{g}(x)^2) + (\overline{g}(x)^2 - 2\overline{g}(x)\mathbb{E}_D[y(x)] + \mathbb{E}_D[y(x)^2])] =$$
$$\mathbb{E}_{xy}[\mathbb{E}_D[(g^D(x)^2 - \overline{g}(x))^2] + (\overline{g}(x) - f(x))^2 - 2\overline{g}(x)\mathbb{E}_D[\epsilon] + 2\mathbb{E}_D[f(x)\epsilon] + \mathbb{E}_D[\epsilon^2]] =$$
$$\mathbb{E}_{xy}[\mathbb{E}_D[(g^D(x)^2 - \overline{g}(x))^2] + (\overline{g}(x) - f(x))^2 + \mathbb{E}_D[\epsilon^2]] =$$
$$\texttt{var} + \texttt{bias} + \sigma^2$$

Demostrando la igualdad dada.

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 E_{in} mayor de 0.008?.

   Solución En las condiciones del Ejercicio que nos indican tenemos que

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

   Buscamos que el valor esperado de E_{in} sea mayor que 0.008, bajo los valores de $\sigma = 0.1$ y $d = 8$, esto no lleva a buscar un tamaño muestral N que verifique:

$$\begin{gathered} 0.008 < 0.01(1 - \frac{9}{N}) \rightarrow \\ 0.008 < 0.01 - \frac{0.09}{N} \rightarrow \\ -0.002 < -\frac{0.09}{N} \rightarrow \\ 0.002 > \frac{0.09}{N} \rightarrow \\ N > 45 \end{gathered}$$

   Es decir, el valor más pequeño para que el tamaño muestral proporcione un valor esperado de E_{in} mayor de 0.008 será N=46.

7. En regresión logística mostrar que

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

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

   Solución
   En regresión logística tenemos

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

   Hagamos el gradiente para verificar la primera igual que nos piden:

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

   (En la segunda igualdad hemos multiplicado y dividido por $e^{-y_n\mathbf{w}^T\mathbf{x}_n}$)

   Ahora vemos la segunda igualdad. Sabemos que $\sigma(s) = \frac{e^s}{1+e^s}$, luego:
   $\sigma(-y_n\mathbf{w}^T\mathbf{x}_n) = \frac{e^{-y_n\mathbf{w}^T\mathbf{x}_n}}{1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}} = \frac{1}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$, luego finalmente se demuestra la segunda igualdad que nos piden:

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

Vamos a argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.
Sabemos que

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

Caso 1: Si tenemos un dato $\mathbf{x}_n$ mal clasificado, tendremos que y_n y $\mathbf{w}^T\mathbf{x}_n$ tienen signo opuesto, luego $y_n\mathbf{w}^T\mathbf{x}_n$ será negativo, quedándonos una exponencial negativa. Esto nos lleva a que el valor $1 + e^{y_n\mathbf{w}^T\mathbf{x}_n}$ será muy cercano a 1.
Caso 2:Si tenemos un dato $\mathbf{x}_n$ bien clasificado, tendremos que y_n y $\mathbf{w}^T\mathbf{x}_n$ tienen el mismo signo, luego $y_n\mathbf{w}^T\mathbf{x}_n$ será positivo, quedándonos una exponencial positiva. Esto nos lleva a que el valor $1 + e^{y_n\mathbf{w}^T\mathbf{x}_n}$ será mucho más grande que en el caso 1. Luego el valor del gradiente será menor que el caso 1 concluyendo que un dato bien clasificado contribuye menos al gradiente que uno mal clasificado.

8. Definamos 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 $\nu = 1$.
Solución
Vamos ir analizando los dos casos para $e_n(\mathbf{w})$ posibles:
Si $e_n(\mathbf{w}) = 0$, es porque $-y_n\mathbf{w}^T\mathbf{x}_n < 0$. Esto quiere decir que x_n está bien clasificado. Tendremos $\nabla e_n(\mathbf{w}) = 0$.
Si $e_n(\mathbf{w}) = -y_n\mathbf{w}^T\mathbf{x}_n$, es porque y_n y $\mathbf{w}^T\mathbf{x}_n$ tienen signo opuesto . Esto quiere decir que x_n no está bien clasificado. Tendremos $\nabla e_n(\mathbf{w}) = -y_n\mathbf{x}_n$.
Una vez que ya tenemos el gradiente para cada situación, podemos afirmar que:
Si x_n está bien clasificado tendremos: $w(t+1) = w(t) - 0$.
Si x_n está mal clasificado tendremos: $w(t+1) = w(t) - (-y_n\mathbf{x}_n) = w(t) + y_n\mathbf{x}_n$ (estamos trabajando con una tasa de aprendizaje de valor 1).
Y de este modo ya se ve que la interpretación es la misma que con el algoritmo PLA.

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.
   Solución
   En este caso, en general, se espera que el ruido determinista suba, ya que para cualquier $g \in \mathcal{H}$ será más difícil aproximar a f (decrementará el ruido estocástico). Recordemos que $E_{out} = \sigma^2 + \texttt{var} + \texttt{bias}$

   (b) Suponer que f es fija y decrementamos la complejidad de $\mathcal{H}$
   Solución
   Si hacemos que la complejidad de $\mathcal{H}$ sea menor, se incrementará el ruido determinista (bias) sin embargo el ruido estocástico (var) decrementará.

   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)

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

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

que define relaciones entre las w_i (La matriz Γ_i se denomina regularizador de Tikhonov)

   (a) Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$
   Solución
   Consideramos el vector $\mathbf{w}$ con Q+1 componentes.
   En este caso podemos tomar Γ como la matriz identidad de dimensión (n)x$(Q+1)$.
   De este modo tendríamos

$$\mathbf{w^t}\Gamma^{\mathrm{T}}\Gamma\mathbf{w} \leq C \to \mathbf{w^t}\mathbf{w} \leq C \to \sum_{q=0}^{Q} w_q^2 \leq C$$

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

Solución

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 X 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)