Aprendizaje Autom
primeatico. Cuestiones optativas 2

Jacinto Carrasco Castillo

27 de abril de 2016

0.1. Matrices y optimización

Cuestión 1 (Multiplicadores de Lagrange)**.** Lagrange propuso una técnica para resolver el siguiente problema de optimización:

$$\max_{x,y} g(x,y)$$

$$\text{Sujeto a } f(x,y) = 0$$

Es decir, buscar el máximo de la función g en un recinto del plano $x - y$ definido por los valores nulos de la función f. La solución es transformar este problema de optimización con restricciones en un problema de optimización sin restricciones y resolver este último derivando e igualando a cero. Para ello construye una nueva función denominada lagrangiana que se define como

$$\mathcal{L}(x,y,\lambda) = g(x,y) - \lambda f(x,y)$$

siendo λ una constante y prueba que la solución de óptimo de $\mathcal{L}$ es la misma que la del problema inicial. Por ello para obtener dicha solución sólo hay que calcular la solución del sistema de ecuaciones dado por $\nabla_{x,y,\lambda}\mathcal{L}(x,y,\lambda) = 0$. En el caso de que exista más de una restricción en igualdad cada una de ellas se añade a la lagrangiana de la misma manera pero con un λ diferente.

$$\mathcal{L}(x,y,\lambda_1,\ldots,\lambda_n) = g(x,y) - \sum_{i=1}^{n} \lambda_i f_i(x,y)$$

Resolver el siguiente problema:

- La distancia entre dos curvas en el plano está dada por el mínimo de la expresión $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ donde (x_1, y_1) está sobre una de las curvas y (x_2, y_2) está sobre la otra. Calcular la distancia entre la línea $x + y = 4$ y la elipse $x^2 + 2y^2 = 1$.

En el caso de que algunas de las condiciones de restricción estén definidas en términos de desigualdad ($<$, $\leq$, etc.), entonces las condiciones para que la solución del problema sin restricción coincida con la solución del problema con restricciones cambian respecto del caso lagrangiano, dichas condiciones se denominan las condiciones de Karush-Kuhn-Tucker.

Solución 1. Minimizaremos $g(x_1, y_1, x_2, y_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ sujeto a $f_1(x, y) = x + y - 4 = 0$, $f_2(x, y) = x^2 + 2y^2 - 1 = 0$. Definimos la función lagrangiana asociada al problema:

$$\mathcal{L}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = g(x_1, y_1, x_2, y_2) + \lambda_1 f_1(x_1, y_1) + \lambda_2 f_2(x_2, y_2)$$

Ahora calculamos su gradiente $\nabla_{x_1, y_1, x_2, y_2, \lambda_1, \lambda_2}$ y lo igualamos a 0:

$$\frac{\partial \mathcal{L}}{\partial x_1}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = \frac{2(x_1 - x_2)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - \lambda_1 = 0 \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial x_2}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = \frac{-2(x_1 - x_2)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - 2\lambda_2 x_2 = 0 \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial y_1}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = \frac{2(y_1 - y_2)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - \lambda_1 = 0 \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial y_2}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = \frac{-2(y_1 - y_2)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - 4\lambda_2 y_2 = 0 \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = -f_1(x_1, y_1) = 0 \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = -f_2(x_2, y_2) = 0 \tag{6}$$

Si igualamos (1) y (2) nos queda:

$$x_1 - x_2 = y_1 - y_2 \tag{7}$$

Si igualamos (3) y (4), usando (7):

$$x_2 = 2y_2 \tag{8}$$

Ahora, sustituyendo en (6), podemos despejar y_2 o x_2:

$$x_2^2 + 2y_2^2 = 1;\ 6y_2^2 = 1;\ y_2 = \pm\frac{1}{\sqrt{6}}; x_2 = \pm\frac{2}{\sqrt{6}} \tag{9}$$

Si despejamos x_1 en (5) y (7) e igualamos, llegamos a:

$$2y_1 = 4 - y_2; y_1 = 2 - \frac{y_2}{2} \tag{10}$$

Y finalmente de (10) y (5) se deduce x_1:

$$x_1 = 4 - y_1 = 4 - 2 + \frac{y_2}{2} = 2 + \frac{y_2}{2} \tag{11}$$

Por tanto, los óptimos locales de esta función están en

$$\mathrm{X}_1 = \begin{pmatrix} x_1 &=& 2 + \frac{1}{2\sqrt{6}} \\ y_1 &=& 2 - \frac{1}{2\sqrt{6}} \\ x_2 &=& \frac{2}{\sqrt{6}} \\ y_2 &=& \frac{1}{\sqrt{6}} \end{pmatrix}, \mathrm{X}_2 = \begin{pmatrix} x_1 &=& 2 - \frac{1}{2\sqrt{6}} \\ y_1 &=& 2 + \frac{1}{2\sqrt{6}} \\ x_2 &=& -\frac{2}{\sqrt{6}} \\ y_2 &=& -\frac{1}{\sqrt{6}} \end{pmatrix}$$

Si evaluamos la función $g(x_1, y_1, x_2, y_2)$ en estos cuatro puntos, vemos que la solución al problema es X_1 y la distancia mínima es aproximadamente 1,9624.

Cuestión 2. La programación lineal es una técnica de optimización que busca el óptimo (máximo o mínimo) de una función lineal en una región de valores definida por un sistema de ecuaciones en desigualdad. En concreto,

$$\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x}$$

$$\text{Sujeto a } A\mathbf{x} \leq \mathbf{b}$$

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

a Argumentar que el problema de programación lineal que encontró en el apartado anterior y el problema de optimización del ejercicio 3 de la sección MINIMIZACIÓN ITERATIVA son equivalentes.

Solución 2. Partimos de la solución al apartado previo. Para llegar a esta solución he realizado el siguiente razonamiento: Supongamos $\mathbf{x}, \mathbf{w}$ de dimensión k. Sea ξ el vector de dimensión N con los ξ_i. Llamamos $\mathrm{X} = \begin{pmatrix} \mathbf{w} \\ \xi \end{pmatrix}$. Entonces, hallar $\min_{\mathbf{x},\xi} \sum_{i=0}^{N} \xi_i$ equivale a hallar $\min_{\mathrm{X}}(0, \overset{k)}{\ldots}, 0, 1, \overset{N)}{\ldots}, 1)^T\mathrm{X}$.

Ahora resulta sencillo ver qué transformación hay que hacerle a la restricción del sistema de ecuaciones de la programación lineal:

$$\begin{pmatrix} -y_1x_1^{(1)} & \dots & -y_1x_1^{(k)} & -1 & 0 & \overset{N)}{\dots} & 0 \\ -y_2x_2^{(1)} & \dots & -y_2x_2^{(k)} & 0 & -1 & \overset{N)}{\dots} & 0 \\ \vdots & \ddots & \vdots & \vdots & & \ddots & \vdots \\ -y_Nx_N^{(1)} & \dots & -y_Nx_N^{(k)} & 0 & 0 & \overset{N)}{\dots} & -1 \\ 0 & \dots & 0 & -1 & 0 & \overset{N)}{\dots} & 0 \\ \vdots & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \overset{N)}{\dots} & -1 \end{pmatrix} \begin{pmatrix} \mathbf{w}^{(1)} \\ \mathbf{w}^{(2)} \\ \vdots \\ \mathbf{w}^{(k)} \\ \xi_1 \\ \vdots \\ \xi_N \end{pmatrix} \leq \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Continuamos ahora con el ejercicio propuesto. Supongamos que tenemos la solución a ambos problemas. Para el apartado anterior, la solución sería $\mathrm{X} = \begin{pmatrix} \mathbf{w} \\ \xi \end{pmatrix}$ de manera que $\sum_{i=0}^{N} \xi_i$ es mínimo. Recordemos que **xi** era el vector de valores de holgura necesarios para que la clasificación fuese correcta, esto es, $y_n(\mathbf{w}^T x_n) \geq 1 - \xi_n \Rightarrow$ Si tenemos $\sum_{i=0}^{N} \xi_i \geq \sum_{i=0}^{N} 1 - y_n(\mathbf{w}^T x_n)$, minimizando $\sum_{i=0}^{N} 1 - y_n(\mathbf{w}^T x_n)$ (objetivo planteado en el problema 3 de la sección de minimización iterativa) podremos minimizar $\sum_{i=0}^{N} \xi_i$, pues sólo tiene como razón para la suma no sea igual que $\xi_i \geq 0$.

0.2. Regresión

Cuestión 3. Consideremos que los datos están generados por una función con ruido $y = \mathbf{w}^{*T}\mathbf{x} + \varepsilon$, donde ε es un término de ruido con media cero y varianza σ^2, que está generado de forma independiente para cada muestra $(\mathbf{x}, y)$. Por tanto el error esperado del mejor ajuste posible a esta función es σ^2.

Supongamos una muestra de datos $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ donde el ruido en cada y_n se nota como ε_n y sea $\varepsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_N)^T$; asumimos que $\mathrm{X^T X}$ es invertible. Seguir los pasos que se muestran a continuación y mostrar que el error esperado (i.e. error esperado de entrenamiento) de regresión lineal respecto de $\mathcal{D}$ está dado por

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

a Mostrar que la estimación de $\mathbf{y}$ está dada por $\hat{\mathbf{y}} = \mathrm{X}\mathbf{w}^\star + H\varepsilon$

b Mostrar que el vector de errores dentro de la muestra $\hat{\mathbf{y}} - \mathbf{y}$ puede expresarse como el producto de una matriz por ε ¿Cuál es la matriz?

c Expresar $E_{in}(\mathbf{w}_{lin})$ en función de ε usando el apartado 2c y simplificar la expresión usando el ejercicio 2b.

d Probar que $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$ usando el apartado anterior y la independencia de los errores, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$. (Ayuda: Tener en cuenta la suma de los elementos de la diagonal de una matriz. Además el apartado 2d también es relevante)

e Para analizar el error esperado fuera de la muestra, vamos a considerar un caso que es fácil de analizar. Consideremos un conjunto de test $\mathcal{D}_{test} = \{(\mathbf{x}_1, y\prime_1), \ldots, (\mathbf{x}_N, y\prime_N)\}$ que comparte los mismas entradas que $\mathcal{D}$ pero con términos de ruido de valor diferente. Notemos el ruido en $y\prime_n$ como $\varepsilon\prime_n$ y sea $\varepsilon\prime = [\varepsilon\prime_1, \varepsilon\prime_2, \ldots, \varepsilon\prime_N]^T$. Definir $E_{test}(\mathbf{w}_{lin})$ como el error cuadrático medio sobre $\mathcal{D}_{test}$.

   a) Probar que $\mathbb{E}_{\mathcal{D},\epsilon\prime}[E_{test}(\mathbf{w}_{lin})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$ Este error de test especial, E_{test}, es un caso muy restrictivo del caso general de error fuera de la muestra.

Solución 3. Notamos que $H = X(X^TX)^{-1}X^T$, donde $X \in \mathcal{M}_{N\times(d+1)}$ y X^TX es invertible por el ejercicio 2 de la sección de matrices y optimización.

a La estimación de $\mathbf{y}$ la realizaremos como lo hemos hecho hasta ahora. Partimos del error dentro de la muestra para una función solución lineal $\mathbf{w}$:

$$E_{in}(\mathbf{w}) = \|\mathbf{y} - \mathbf{w}^T\mathbf{w}\|$$

De aquí deducimos haciendo el gradiente en $\mathbf{w}$ que el $\mathbf{w}$ que minimiza este error, con X^TX invertible es $\mathbf{w}\prime = (X^TX)^{-1}X^T\mathbf{y}$. La estimación que haríamos para $\mathbf{y}$ se haría de esta forma, llegando a

$$\hat{\mathbf{y}} = X\mathbf{w}\prime = X(X^TX)^{-1}X^T\mathbf{y}$$

Ahora usamos que sabemos que $\mathbf{y} = X\mathbf{w}^\star + \varepsilon$, luego nos queda:

$$\hat{\mathbf{y}} = X(X^TX)^{-1}X^T(X\mathbf{w}^\star + \varepsilon)$$

Aplicando la propiedad distributiva y sustituyendo H por su valor antes mencionado tenemos

$$\hat{\mathbf{y}} = X\mathbf{w}^\star + H\varepsilon)$$

b Ahora simplemente tenemos que restar $\hat{\mathbf{y}}$ y $\mathbf{y}$ y sustituir cada valor por lo que hemos calculado previamente.

$$\hat{\mathbf{y}} - \mathbf{y} = X\mathbf{w}^\star + H\varepsilon - X\mathbf{w}^\star - \varepsilon = (H - I)\varepsilon$$

c $E_{in}(\mathbf{w}_{lin}) = \frac{1}{N}\|\hat{\mathbf{y}} - \mathbf{y}\|$ Calculamos a partir del punto anterior:

$$\frac{1}{N}\|\hat{\mathbf{y}} - \mathbf{y}\| = \frac{1}{N}((H - I)\varepsilon)^T(H - I)\varepsilon =$$

$$= \frac{1}{N}\varepsilon^T(H - I)^T(H - I)\varepsilon$$

Por ser H simétrica, al restarle la matriz identidad (que sólo tiene elementos distintos de 0 en la diagonal) la matriz resultante seguirá siendo simétrica, luego $(H - I)^T(H - I) = (H - I)^2 = (-1)^2(I - H) = (I - H)$, por ser $(I - H)^K = I - H$ para todo K. Entonces:

$$E_{in}(\mathbf{w}_{lin}) = \frac{1}{N}\varepsilon^T(I - H)\varepsilon$$

d Usamos en primer lugar que la linealidad de la esperanza matemática:

$$\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\varepsilon^T\varepsilon] - \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\varepsilon^T H\varepsilon]$$

Volvemos a aplicar la linealidad en el primer sumando, usando que $\varepsilon^\varepsilon = \sum_{i=0}^{N}\varepsilon_i^2$ y que ε sigue una distribución con media 0 y varianza σ^2

$$\frac{1}{N}\mathbb{E}_{\mathcal{D}}[\varepsilon^T\varepsilon] = \frac{N\sigma^2}{N} = \sigma^2$$

Vamos ahora con el otro término. Usaremos que la esperanza del error es 0, y además que estos errores son independientes los unos de los otros, con lo que $\mathbb{E}_{\mathcal{D}}[\varepsilon_i\varepsilon_j] = \mathbb{E}_{\mathcal{D}}[\varepsilon_i]\mathbb{E}_{\mathcal{D}}[\varepsilon_j] = 0$ para $i \neq j$. Entonces, al calcular la esperanza de $\varepsilon^T H\varepsilon$, nos quedarán únicamente la suma de los elementos de su diagonal (su traza, donde hemos aplicado otra vez la linealidad de la esperanza), multiplicado de nuevo por σ^2, y usamos el ejercicio anterior y tenemos que su traza es $d + 1$, con lo que

$$\mathbb{E}_{\mathcal{D}}[\varepsilon^T H\varepsilon] = \frac{(d + 1) - \sigma^2}{N}$$

Finalmente, obtenemos

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

Cuestión 4. En regresión lineal, el error fuera de la muestra está dado por

$$E_{out}(h) = \mathbb{E}[(h(\mathbf{x}) - y)^2]$$

Mostrar que entre todas las posibles hipótesis, la que minimiza E_{out} está dada por

$$h^\star(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$$

La función $h^\star$ puede considerarse una función determinista, en cuyo caso podemos escribir $y = h^\star(\mathbf{x}) + \varepsilon(\mathbf{x})$, donde $\varepsilon(\mathbf{x})$ es una varible de ruido independiente. Mostrar que $\varepsilon(\mathbf{x})$ tiene valor esperado 0.

Solución 4.

0.3. Regresión Logística

Cuestión 5. Supongamos que queremos predecir una función objetivo con error (i.e. estocástica) $P(y|\mathbf{x})$ a partir de muestras etiquetadas con valores ± 1 y de funciones hipótesis que notamos por h

a Escribir la función de máxima verosimilitud de una muestra de tamaño N

b Mostrar que la estimación de máxima verosimilitud se reduce a la tarea de encontrar la función h que minimiza

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

c Para el caso $h(x) = \sigma(\mathbf{w}^T\mathbf{x})$ mostrar que minimizar el error de la muestra en el apartado anterior es equivalente a minimizar el error muestral

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

Nota: dadas dos distribuciones de probabilidad $\{p, 1-p\}$ y $\{q, 1-q\}$ de variables aleatorias binarias, la entropía cruzada para estas distribuciones se define en teoría de la información por la expresión

$$p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q}$$

El error de la muestra en el apartado.b corresponde a una medida de error de entropía cruzada de los datos $(\mathbf{x}_n, y_n)$ con $p = [y_n = +1]$ y $q = h(\mathbf{x}_n)$.

Solución 5.

Cuestión 6. Consideremos el caso de la verificación de la huella digital (ver transparencias de clase). Tras aprender con un modelo de regresión logística a partir de datos obtenemos una función una hipótesis final

$$g(x) = \mathbb{P}[y = +1|\mathbf{x}]$$

que representa la estimación de la probabilidad de que $y = +1$. Suponga que la matriz de coste está dada por

		Verdadera Clasificación	
		+1 (persona correcta)	-1 (intruso)
decisión	+1	0	c_a
decisión	-1	c_r	0

Para una nueva persona con huella digital $\mathbf{x}$, calculamos $g(\mathbf{x})$ y tenemos que decidir si aceptar o rechazar a la persona (i.e. tenemos que usar una decisión 1/0). Por tanto aceptaremos si $g(\mathbf{x}) \geq \kappa$, donde κ es un umbral.

a Definir la función de costo (aceptar) como el costo esperado si se acepta la persona. Definir de forma similar el costo (rechazo). Mostrar que

$$\begin{aligned} \text{costo(aceptar)} &= (1 - g(\mathbf{x}))c_a \\ \text{costo(rechazar)} &= g(\mathbf{x})c_r \end{aligned}$$

b Usar el apartado anterior para derivar una condición sobre $g(x)$ para aceptar la persona y mostrar que

$$\kappa = \frac{c_a}{c_a + c_r}$$

c Usar las matrices de costo para la aplicación del supermercado y la CIA (transparencias de clase) para calcular el umbral κ para cada una de las dos clases. Dar alguna interpretación del umbral obtenido.

Solución 6.

0.4. Regularización y selección de modelos

Cuestión 7. Dentro de los modelos de regresión lineal se han hecho numerosos intentos de calcular el número efectivo de parámetros en un modelo. Tres de las posibilidade son

a $d_{eff}(\lambda) = 2trace(H(\lambda)) - trace(H^2(\lambda))$

b $d_{eff}(\lambda) = trace(H(\lambda))$

c $d_{eff}(\lambda) = trace(H^2(\lambda))$

donde $H(\lambda) = Z(Z^T Z + \lambda I)^{-1} Z^T$ y Z es la matriz de datos transformados. Para obtener d_{eff} debemos primero calcular $H(\lambda))$ igual que cuando hacemos regresión. Entonces podemos usar de forma heurística d_{eff} en lugar de d_{VC} en la cota de generalización.

1. Cuando $\lambda = 0$ mostrar que para las tres elecciones, $d_{eff} = \bar{d}+1$, donde $\bar{d}$ es la dimensión en el espacio $\mathcal{Z}$.

2. Cuando $\lambda > 0$ mostrar que $0 \leq d_{eff} \leq \bar{d} + 1$ y d_{eff} es decreciente en λ para las tres opciones. (Ayuda: usar SVD)

Solución 7.