Aprendizaje Automático: Bonus 2

Anabel Gómez Ríos

27 de abril de 2016

0.1. Matrices y optimización. Problema 3.c

Pregunta 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.

Vamos a definir por tanto para nuestro caso el operador $\mathcal{L}$. Necesitamos dos parejas de puntos, una sujeta a pertenecer a la recta $x_1 + y_1 = 4$ y otra sujeta a pertenecer a la elipse $x_2^2 + 2y_2^2 = 1$ (necesitamos por tantos dos λ distintos: λ_1, λ_2). Por tanto en nuestro caso tenemos

$$\mathcal{L}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} - \lambda_1(x_1 + y_1 - 4) - \lambda_2(x_2^2 + 2y_2^2 - 1) \tag{1}$$

Como nos dice el enunciado del problema, tenemos ahora que calcular el gradiente con respecto a todas las variables (las dos parejas de puntos y los dos parámetros) y resolver el sistema resultante de igualar este gradiente a cero. Calculamos el gradiente, derivando $\mathcal{L}$ con respecto a todas las variables y dando como resultado un vector de 6 componentes:

$$\nabla_{x_1,y_1,x_2,y_2,\lambda_1,\lambda_2}\mathcal{L}(x_1, y_1, x_2, y_2, \lambda_1, \lambda_2) =$$

$$= \left(\frac{2(x_1 - x_2)}{2\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - \lambda_1, \frac{2(y_1 - y_2)}{2\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - \lambda_1,\right.$$

$$\frac{2(x_1 - x_2)}{2\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - 2\lambda_2 x_2, \frac{2(y_1 - y_2)}{2\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} - 4\lambda_2 y_2,$$

$$\left. -x_1 - y_1 + 4, -x_2^2 - 2y_2^2 + 1\right)$$

Por simplicidad, vamos a notar a la raíz cuadrada $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ como A. Nos queda, simplificando un poco, el siguiente sistema de 6 ecuaciones con 6 incógnitas:

$$\frac{x_1 - x_2}{A} - \lambda_1 = 0$$

$$\frac{y_1 - y_2}{A} - \lambda_1 = 0$$

$$\frac{-x_1 + x_2}{A} - 2\lambda_2 x_2 = 0$$
$$\frac{-y_1 + y_2}{A} - 4\lambda_2 y_2 = 0$$
$$-x_1 - y_1 + 4 = 0$$
$$-x_2^2 - 2y_2^2 + 1 = 0$$

De las dos primeras ecuaciones, vamos a despejar λ_1 y a igualar, y nos queda por tanto $\frac{x_1 - x_2}{A} = \frac{y_1 - y_2}{A} \Rightarrow x_1 - x_2 = y_1 - y_2 \Rightarrow x_1 = x_2 + y_1 - y_2$

De las dos siguientes ecuaciones vamos a despejar los numeradores en ambas, con lo que nos queda lo siguiente: $x_1 - x_2 = -2\lambda_2 x_2 A$ en la primera y $y_1 - y_2 = -4\lambda_2 y_2 A$ en la segunda. Como acabamos de obtener de las dos primeras ecuaciones que $x_1 - x_2 = y_1 - y_2$ podemos igualar los segundos miembros y dividiendo por 2 nos queda $\lambda_2 x_2 A = 2\lambda_2 y_2 A \Rightarrow x_2 = 2y_2$.

Ahora en la última ecuación vamos a utilizar que $x_2 = 2y_2$ y a sustituir, de forma que tenemos $-x_2^2 - 2y_2^2 + 1 = 0 \Rightarrow -4y_2^2 - 2y_2^2 + 1 = 0 \Rightarrow 6y_2^2 = 1 \Rightarrow y_2^2 = \frac{1}{6} \Rightarrow y = \pm\frac{1}{\sqrt{6}}$ con lo que obtenemos dos valores para y_2.

Ahora que tenemos y_2 vamos a la penúltima ecuación y a utilizarlo junto con que $x_1 = x_2 + y_1 - y_2$ y que $x_2 = 2y_2$: $-x_1 - y_1 + 4 = 0 \Rightarrow -x_2 - y_1 + y_2 - y_1 + 4 = 0 \Rightarrow -2y_2 + y_2 - 2y_1 + 4 = 0 \Rightarrow 2y_1 + y_2 = 4 \Rightarrow y_1 = 2 - \frac{y_2}{2}$. Con esto obtenemos el valor de de y_1, que serán también dos valores, dependiendo de si cogemos el signo negativo o el positivo de y_2: $y_1 = 2 - (\pm\frac{1}{2\sqrt{6}})$.

Como tenemos y_2 podemos obtener también $x_2 = 2y_2 = \pm\frac{2}{\sqrt{6}}$.
Una vez tenemos x_2, y_1, y_2 podemos obtener x_1, $x_1 = x_2 + y_1 - y_2 = y_1 + y_2 = 2 - (\pm\frac{1}{2\sqrt{6}}) \pm \frac{1}{\sqrt{6}}$. Tenemos que tener en cuenta que el signo de estos tres (x_1, x_2, y_1), cuando hemos puesto $\pm$ hasta ahora, es siempre el mismo puesto que todos dependen del signo de y_2 (es decir, para y_2 positivo todos los $\pm$ serán un $+$ y si es negativo todos serán un $-$).

Cabe destacar que una vez tenemos los dos puntos obtener λ_1 y λ_2 es fácil, pero no es necesario para nuestro problema ya que hemos llegado a una solución sin la necesidad de calcularlos, es decir, nosotros lo que necesitamos

es justo estos dos puntos. Hemos llegado por tanto a dos soluciones distintas:

$$(x_1, y_1, x_2, y_2) = (2-\frac{1}{2\sqrt{6}}+\frac{1}{\sqrt{6}}, 2-\frac{1}{2\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}) = (2+\frac{1}{2\sqrt{6}}, 2-\frac{1}{2\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}})$$

$$(x_1, y_1, x_2, y_2) = (2+\frac{1}{2\sqrt{6}}-\frac{1}{\sqrt{6}}, 2+\frac{1}{2\sqrt{6}}, -\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}) = (2-\frac{1}{2\sqrt{6}}, 2+\frac{1}{2\sqrt{6}}, -\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}})$$

Vamos a ver cuál de estas dos soluciones nos da una distancia más pequeña, puesto que lo que queríamos calcular era la mínima distancia entre la recta y la elipse.
Para la primera solución, tenemos:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = 1{,}435341$$

y para la segunda solución tenemos:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = 4{,}474721$$

Con lo que vemos que la distancia más pequeña es la de la primera solución y por tanto la solución a nuestro problema de minimización es

$$(x_1, y_1, x_2, y_2) = (2 + \frac{1}{2\sqrt{6}}, 2 - \frac{1}{2\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}})$$

0.2. Regresión Logística. Problema 3

Pregunta 2. 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.

a) g es la probabilidad de que, ante una huella digital $\mathbf{x}$, se le permita a la persona entrar (es decir, se dé la salida $y_n = +1$). Como vemos en la matriz de costos, el coste de aceptar a una persona cuando no deberíamos haberla aceptado es c_a y el coste de rechazar a una persona cuando no deberíamos haberla rechazado es c_r. El costo esperado por tanto de aceptar a una persona es el coste de aceptarla y equivocarnos (aceptar a un intruso) por la probabilidad de que la persona efectivamente sea un intruso, es decir, $1 - g(x)$, con lo que el coste estimado de aceptar es $c_a(1 - g(x))$.
Volviendo de nuevo a la matriz de costos, vemos que el coste de rechazar a una persona cuando deberíamos haberla aceptado es c_r. Entonces, el coste esperado de rechazar a una persona es el coste de rechazarla y equivocarnos por la probabilidad de que efectivamente la huella fuera correcta, que es exactamente $g(x)$, es decir, coste(rechazo)=$c_r g(x)$.
Con esto obtenemos como vemos lo que se nos pedía.

b) De forma intuitiva, podemos pensar que si queremos que se acepte a poca gente, tendremos que subir el umbral para el que se acepta a alguien y si queremos aceptar a mucha gente, bajarlo. Este umbral es exactamente κ. Como nos pide una condición sobre la $g(x)$ vamos a pensar primero sobre ella. Esta probabilidad (que es la probabilidad de aceptar a una persona) nos interesa que sea mayor cuando el coste de aceptar a una persona

y equivocarnos (c_a) sea bajo y cuando el coste de rechazar a alguien y equivocarnos (c_r) sea alto, ya que estamos minimizando así el coste total (habrá menos probabilidad de rechazar a alguien). Del mismo modo, nos interesa que sea menor cuando la probabilidad de equivocarnos aceptando a alguien(c_a) sea alta y la probabilidad de equivocarnos rechazando a alguien (c_r) sea baja, puesto que estaremos permitiendo la entrada de menos personas y estaremos minimizando así el coste total de nuevo. Sin embargo nosotros no podemos tocar esta probabilidad, lo que sí podemos hacer es cambiar el umbral a partir del cual vamos a aceptar a alguien, como hemos comentado previamente, de forma que éste subirá cuando queramos aceptar a pocas personas y bajará cuando queramos aceptar a muchas. Tenemos que razonar por tanto de manera inversa que como lo hemos hecho con $g(x)$, ya que se mueven al revés. De esta forma tenemos que si c_a es bajo o c_r es alto, queremos que κ sea alto y si c_a es alto o c_r es bajo, que κ suba.
Por tanto, si ponemos κ como un cociente y queremos que sea pequeño cuando c_a sea pequeño y c_r grande, tendremos que poner c_a en el numerador y c_r en el denominador. De las otras condiciones deducimos lo mismo, ya que si queremos que κ sea alto cuando c_a es alto y c_r bajo, tenemos que poner c_a en el numerador y c_r en el denominador. Tenemos ahora mismo $\kappa = \frac{c_a}{c_r}$. Sin embargo tenemos que tener en cuenta que κ está acotando por debajo una probabilidad, que será siempre un número entre 0 y 1, por lo que κ tendrá que estar también entre 0 y 1. Esto se consigue sumándole al denominador el numerador: $\kappa = \frac{c_a}{c_r+c_a}$, ya que así el denominador es más grande siempre (asumiendo que los costes serán siempre positivos) y valdrá como mucho 1 y se tenderá a 1 cuanto más grande sea c_a, mientras que si c_r es más grande, tenderá a 0, que es justo lo que queríamos.

c) Vamos a calcular κ para las dos aplicaciones. Empezamos con la del supermercado, donde $c_a = 1$ y $c_r = 10$, entonces $\kappa = \frac{1}{10+1} = 0{,}0909$. En el caso de la CIA, $c_a = 1000$ y $c_r = 1$ con lo que $\kappa = \frac{1000}{1000+1} = 0{,}999$. Como vemos, en el supermercado el umbral es mucho más bajo, ya que no nos importa aceptar a una persona que no sea cliente habitual tanto como no aceptar a una que sí lo sea, por lo que el umbral de la probabilidad está cercano a 0, mientras que en el de la CIA es extremadamente importante no aceptar a nadie que no sea miembro, por lo que el umbral está cerca de 1, en el sentido en el que hemos comentado en el apartado anterior.