Problemas Convexos

Carlos Malanche

20 de marzo de 2018

1. Máxima verosimilitud (*Maximum Likelihood*)

Podemos ver el problema de una manera alterna y un tanto complicada, es una manera de ver las cosas basados en probabilidad.

Como ya lo habíamos dicho antes, podemos suponer que nuestra información fue generada por una parte determinística más una parte de ruido

$$y = f(\boldsymbol{x}) + \epsilon \tag{1}$$

Con una restricción más, supongamos que la función determinística es una función lineal, tal y como lo hicimos con mínimos cuadrados

$$y = \boldsymbol{x}^T \boldsymbol{c} + \epsilon \tag{2}$$

Dado un $\underset{\sim}{c}$ fijo, podemos preguntarnos cuál es la probabilidad de haber obtenido las mediciones y.

$$P(y|\boldsymbol{x}, \boldsymbol{c}) = \mathcal{N}(y|\boldsymbol{x}^T\boldsymbol{c}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\boldsymbol{x}^T\boldsymbol{c})^2}{2\sigma^2}} \tag{3}$$

En el caso de $\boldsymbol{y}$ un vector, la expresión es un tanto más complicada

$$P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{c}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{c}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\Big[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) \Big] \tag{4}$$

La idea de *maximum likelihood* es encontrar el vector de parámetros $\underset{\sim}{c}$ que maximiza la probabilidad de haber obtenido el vector de *outcomes* $\boldsymbol{y}$.

Para resolver esto: ¿La probabilidad de obtener $\boldsymbol{y}$ dado $\boldsymbol{X}$ como función de $\boldsymbol{c}$ qué forma tiene? ¿Es un problema convexo? ¡No! y tampoco es cóncavo. Primero vamos a arreglar lo primero. Hay una forma cuadrática en el argumento de la exponencial, la cual ya sabemos será convexa (excepto por el signo menos...). Vamos a intentar sacarla aplicando un logaritmo a la probabilidad.

$$\log(P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{c})) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) + \text{cnst} \tag{5}$$

En donde hemos decidido abreviar esa constante

$$\text{cnst} = -\frac{1}{2} m \log((2\pi)^m \det(\Sigma))$$

Basta con multiplicar toda la expresión por un menos y deshacernos de la constante (no va a modificar el problema) para movernos a un marco muy similar al que estabamos utilizandoa anteriormente (un problema de minimizar en lugar de maximizar).

$$\arg\min_{\underset{\sim}{c}} \{ \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) \} \tag{6}$$

El problema es casi idéntico al de mínimos cuadrados! En realidad, es una generalización, pues basta con hacer $\Sigma = n\mathbb{I}$ para recuperar los mínimos cuadrados. Hacer esto le quitaría el sentido estadístico a nuestra interpretación así que no lo haremos.

En lugar de ello, todo se simplifica cuando las matriz de covarianza es diagonal, pues implica que cada una de las medidas fue independiente (lo cual es algo relativamente seguro de hacer en la mayoría de los casos). Incluso

podemos asumir que la desviación estándar σ es la misma para cada uno de los parámetros, con lo que obtenemos casi la función de pérdidas cuadrática

$$\arg\min_{\underset{\sim}{c}}\{\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{c})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{c})\} \tag{7}$$

2. Regresión regularizada

Recordemos que el objetivo es encontrar una vector de parámetros (los cuales definen una función) que minimize el error de predicción sobre una serie de datos conocida. Si seguimos denotando con $J_{\hat{f}}(\underset{\sim}{c})$ la función de costo adaptada a nuestra función estimadora $\hat{f}$, definiremos pues la función objetivo como

$$\mathcal{L}(\underset{\sim}{c}) \coloneqq J_{\hat{f}}(\underset{\sim}{c}) \tag{8}$$

se ve un poco redundante pero tiene un propósito inventar esta nueva función. Lo que buscamos es minimizar la función objetivo

$$\arg\min_{\underset{\sim}{c}} \mathcal{L}(\underset{\sim}{c}) \tag{9}$$

Hemos visto que cuando J es la función de pérdidas cuadrática (FPC) la solución es analítica y está dada por las ecuaciones normales. Sin embargo, el problema puede volverse inestable dependiendo del rango de la matriz de observaciones $\boldsymbol{X}$ (columnas casi linealmente independientes convierten en casi-singular a la matriz, y su inversión es inestable). Llegan a salvarnos los **regularizadores**, que son funciones que se añaden a la función objetivo para poner imponer condiciones.

$$\mathcal{L}(\underset{\sim}{c}) \coloneqq J_{\hat{f}}(\underset{\sim}{c}) + \Omega(\underset{\sim}{c}) \tag{10}$$

Esto es todo un campo de investigación, pues no basta con pasarle condiciones a la funció objetivo a diestra y siniestra, no tiene caso hacerlo si no sabemos como resolver el problema. Un ejemplo de un *regularizador* es esta modificación a la función indicadora

$$\Omega(\underset{\sim}{c}) = 1_S \coloneqq \begin{cases} 0, & \text{si } \underset{\sim}{c} \notin S \\ \infty, & \text{o.c.} \end{cases} \tag{11}$$

Basta con que la función de costo no sea infinita para una $\underset{\sim}{c} \in S$ para que el vector solución $\underset{\sim}{c}$ no pueda estar fuera de S.

Por el momento nos vamos a enfocar en regularizadores con un comportamiento un poco menos agresivo (en general, derivables) que de todos modos nos pueden ayudar, en particular a que la matriz de las ecuaciones normales no quede mal condicionada.

2.1. Regularización de Tikhonov

Definamos los regularizadores de Tikhonov como aquellas función objetivo en donde el regularizador Ω tiene la forma

$$\Omega(\underset{\sim}{c}) \coloneqq \|\boldsymbol{\Gamma}\boldsymbol{c}\|^2 \tag{12}$$

Donde $\boldsymbol{\Gamma}$ es una matriz que podemos elegir para sacar características lineales del vector de parámetros. Busquen cuál es el resultado de encontrar el gradiente del problema con regularización de Tikhonov, es muy sencillo.

2.2. Regularización de arista

Caso particular, tomaremos

$$\Omega(\underset{\sim}{c}) \coloneqq \frac{\gamma}{2n}\|\underset{\sim}{c}\|^2 \tag{13}$$

es decir, $\boldsymbol{\Gamma} = (\gamma/2n)^{1/2}\mathbb{I}$. Vamos a repetir el ejercicio de la clase pasada, y vamos a tratar de encontrar un punto óptimo para este caso

$$\nabla\mathcal{L}(\underset{\rightarrow}{c}) = \nabla(\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}\|^2) + \frac{\gamma}{2n}\nabla\|\underset{\rightarrow}{c}\|^2 \tag{14}$$

$$= \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}) + \frac{\gamma}{n}\boldsymbol{c} \tag{15}$$

Ya sólo igualamos a cero y resolvemos para $\boldsymbol{c}$

$$\boldsymbol{c} = (\boldsymbol{X}^T\boldsymbol{X} + \gamma\mathbb{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{16}$$

Para los atentos, el efecto de sumarle ese múltiplo de la identidad a una matriz semidefinida positiva es el de levantar los eigenvalores, con lo que podemos obligar a la matriz completa a ser invertible

Teorema. *Sea $\boldsymbol{G}$ una matriz de Gram generada por la matriz $\boldsymbol{X}$. Entonces los valores propios de $(\boldsymbol{G} + \gamma\mathbb{I})$ se elevan por γ*

Demostración. Consideremos la diagonalización de la matriz de Gram, dada por las matriz de vectores propios $\boldsymbol{V}$ y la matriz diagonal de valores propios $\boldsymbol{P}$.

$$\boldsymbol{G} = \boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V}$$

Con ella, escribimos de nuevo nuestra matriz por invertir

$$\boldsymbol{X}^T\boldsymbol{X} + \gamma\mathbb{I} = \boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V} + \gamma\mathbb{I} = \boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V} + \gamma\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}^T(\boldsymbol{P} + \gamma\mathbb{I})\boldsymbol{V} \tag{17}$$

Nota: Esta demostración se auxilió de saber que las matrices de vectores propios de una matriz de Gram son ortonormales, lo cual se puede ver con la descomposición de valores singulares de su matriz generadora. □

Casi por diversión, de una manera alternativa podemos extraer los valores propios usando el coeficiente de *Rayleigh*

$$\frac{\boldsymbol{v}^T(\boldsymbol{X}^T\boldsymbol{X} + \gamma\mathbb{I})\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} = \frac{\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} + \gamma\frac{\boldsymbol{v}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} = \frac{\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} + \gamma$$

Añadimos el conocimiento de que esa matriz de Gram es semidefinida positiva, lo cual quiere decir que los valores propios son iguales o mayores a cero, con lo que podemos concluír que los valores propios de la matriz por invertir **son almenos** γ.

2.3. Regularización *Lasso*

Otro caso divertido es la regularización de *Lasso* (por sus siglas *least absolute shrinkage and selection operator*), en donde la función de regularización es la norma en L_1, es decir

$$\Omega(\underset{\rightarrow}{c}) = \|\underset{\rightarrow}{c}\|_{L_1}$$

léase, la suma de los parámetros que definen al modelo. Lo curioso de este regularizador es que favorece soluciones con muchas entradas en cero. Piensen en un caso simple de dos dimensiones, y se podrán imaginar por qué pasa esto.

3. Problemas convexos

Nosotros agarramos y minimizamos el problema asumiendo que era convexo pero no lo probamos. Pues bien, definamos lo que es un problema convexo

Definición. *Una función $f : U \to \mathbb{R}$ definida en un conjunto convexo y abierto es convexa si para cualesquiera dos puntos $\boldsymbol{s}_1, \boldsymbol{s}_2 \in U \subset \mathbb{R}^m$ y $\lambda \in [0, 1]$ se cumple que*

$$f(\lambda\boldsymbol{s}_1 + (1 - \lambda)\boldsymbol{s}_2) \leq \lambda f(\boldsymbol{s}_1) + (1 - \lambda)f(\boldsymbol{s}_2)$$

La definición es muy intuitiva, pues nos dice que el resultado de evaluar f en cualquiera de los puntos sobre la recta que une un par de puntos $\underset{\rightarrow}{s}_2, \underset{\rightarrow}{s}_2 \in S$ (donde S es el dominio de la función) es inferior al resultado de evaluar el mismo punto en la línea que une $f(\underset{\rightarrow}{s}_1)$ con $f(\underset{\rightarrow}{s}_2)$.

Para el caso de mínimos cuadrados, se tendría que verificar que la siguiente desigualdad es cierta

$$\frac{1}{2n}\|\boldsymbol{X}(\lambda \boldsymbol{s}_1 + (1-\lambda)\boldsymbol{s}_2) - \boldsymbol{y}\|^2 \leq \lambda \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{s}_1 - \boldsymbol{y}\|^2 + (1-\lambda)\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{s}_2 - \boldsymbol{y}\|^2 \tag{18}$$

Ese ejercicio queda al lector. Vamos a intentar llegar a una definición alterna.

$$\begin{aligned} f(\lambda \boldsymbol{s}_1 + (1-\lambda)\boldsymbol{s}_2) &\leq \lambda f(\boldsymbol{s}_1) + (1-\lambda) f(\boldsymbol{s}_2) \\ f(\lambda(\boldsymbol{s}_1 - \boldsymbol{s}_2) + \boldsymbol{s}_2) &\leq \lambda(f(\boldsymbol{s}_1) - f(\boldsymbol{s}_2)) + f(\boldsymbol{s}_2) \\ \frac{f(\lambda(\boldsymbol{s}_1 - \boldsymbol{s}_2) + \boldsymbol{s}_2) - f(\boldsymbol{s}_2)}{\lambda} &\leq f(\boldsymbol{s}_1) - f(\boldsymbol{s}_2) \end{aligned}$$

Sin pérdida de generalidad, vamos a tomar el límite cuando λ tienda a cero (ya suficiente libertad tenemos con permitir a $\underset{\rightarrow}{s}_1$ y a $\underset{\rightarrow}{s}_2$ ser cualesquiera dos puntos en el dominio)

$$\begin{aligned} \lim_{\lambda \to 0} \frac{f(\lambda(\boldsymbol{s}_1 - \boldsymbol{s}_2) + \boldsymbol{s}_2) - f(\boldsymbol{s}_2)}{\lambda} &\leq f(\boldsymbol{s}_1) - f(\boldsymbol{s}_2) \\ (\boldsymbol{s}_1 - \boldsymbol{s_2})^T \nabla f(\boldsymbol{s}_2) &\leq f(\boldsymbol{s}_1) - f(\boldsymbol{s}_2) \end{aligned}$$

Donde hemos hecho uso de una de las definiciones de derivada direccional (una que no se encuentra normalizada). Haciendo un último despeje, tenemos esta definición alterna para funciones diferenciables:

Definición. *Una función $f : \mathbb{R}^m \to \mathbb{R}$ diferenciable es convexa si para cualesquiera dos puntos $\boldsymbol{s}_1, \boldsymbol{s}_2 \in S \subset \mathbb{R}^m$ se cumple que*

$$f(\boldsymbol{s}_1) - f(\boldsymbol{s}_2) \geq (\boldsymbol{s}_1 - \boldsymbol{s_2})^T \nabla f(\boldsymbol{s}_2)$$