UC | Chile

UC | Chile

Algebra Lineal Aplicada para Ciencia de Datos

Clase 11. Penalización cuadrática

1. Motivación
2. Formulación del problema
3. Solución usando mínimos cuadrados ordinarios
4. Solución usando la descomposición en valores singulares
5. La pseudoinversa
6. Generalizaciones

UC | Chile

Motivación

Motivación

Los **sistemas subdeterminados**

$$Ax = y$$

con $y \in \mathbb{R}^m$ y $A \in \mathbb{R}^{m \times n}$ con $m \leq n$ de **rango completo** tienen **una infinidad de soluciones**

Por lo tanto, es necesario **seleccionar** una solución

Motivación

La clase anterior estudiamos la **solución de mínima norma Euclideana** $x^\star$

Esta resuelve

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|x\|_2^2 \qquad \text{sujeto a} \quad Ax = y$$

y se puede representar como

$$x^\star = A^\top (AA^\top)^{-1} y$$

Por construcción,

$$Ax^\star = y$$

Motivación

Frecuentemente el vector y representa **datos experimentales** sujetos a **perturbaciones**

En este caso, la solución de mínima norma Euclideana es el vector $\widehat{x}$ que resuelve

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|x\|_2^2 \qquad \text{sujeto a} \quad Ax = y + \Delta y$$

donde $\Delta y \in \mathbb{R}^m$ representa esta perturbación

Si Δy es de magnitud pequeña, **¿está $\widehat{x}$ cerca de $x^\star$?**

Motivación. Ejemplo

$$A = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 10^{-9} & 1 + 10^{-9} \end{bmatrix}$$

Para esta matriz se tiene

$$A^\top (AA^\top)^{-1} = \frac{1}{3} \begin{bmatrix} 2 + 10^9 & 10^9 \\ 1 + 2 \cdot 10^9 & 2 \cdot 10^9 \\ -1 + 2 \cdot 10^9 & 0 \end{bmatrix} = \begin{bmatrix} 2/3 & 0 \\ 1/3 & 0 \\ -1/3 & 0 \end{bmatrix} + 10^9 \begin{bmatrix} 1/3 & 1/3 \\ 2/3 & 2/3 \\ 2/3 & 0 \end{bmatrix}.$$

Motivación. Ejemplo

Si la perturbación es

$$\Delta y = 10^{-3} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

entonces

$$\widehat{x} = A^\top (AA^\top)^{-1}(y + \Delta y) = x^\star + 10^{-3} \begin{bmatrix} 2/3 \\ 1/3 \\ -1/3 \end{bmatrix} + 10^{6} \begin{bmatrix} 2/3 \\ 4/3 \\ 2/3 \end{bmatrix}.$$

La **magnitud de la perturbación en la solución** es 10^9 **veces mayor** a **la magnitud de la perturbación en los datos**

Motivación

Cuando esto ocurre, decimos que el sistema es **inestable**

El ejemplo muestra **no es** un problema **numérico** sino que de la **estructura** de A

Para **estabilizar** la solución al problema, estudiamos la **penalización cuadrática**

UC | Chile

Formulación del problema

Formulación del problema

Sea $y \in \mathbb{R}^m$ y $A \in \mathbb{R}^{m \times n}$

No supondremos ninguna relación entre m y n: el sistema lineal

$$Ax = y$$

puede no tener solución, tener una única solución, o una infinidad de soluciones

Buscamos soluciones aproximadas a este sistema que sean estables bajo perturbaciones de y

Formulación del problema

El **residuo** $r(x) = y - Ax$ determina qué tan cerca está x de ser una solución

Buscamos un vector que encuentre un balance entre **reducir la magnitud del residuo** y **penalizar la norma del vector solución**

La **penalización cuadrática** busca un $x^\star$ que resuelva

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|r(x)\|_2^2 + \lambda \|x\|_2^2$$

donde $\lambda > 0$ es un **parámetro de penalización** que controla este balance

Formulación del problema

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|y - Ax\|_2^2 + \lambda \|x\|_2^2$$

Este problema se conoce con distintos nombres

- En **estadística** se conoce como **regresión contraída** o **regresión *ridge***
- En **problemas inversos** se conoce como **regularización Tikhonov**
- En **métodos Bayesianos** este problema permite determinar el **máximo *a posteriori* (MAP)** cuando se usa un *a priori* Gaussiano y la perturbación es Gaussiana

UC | Chile

Solución usando mínimos cuadrados ordinarios

Solución usando mínimos cuadrados ordinarios

Para resolver el problema, primero escribimos

$$\|r(x)\|_2^2 + \lambda\|x\|_2^2 = \left\| \begin{bmatrix} y - Ax \\ \sqrt{\lambda}x \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Ax \\ \sqrt{\lambda}x \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \sqrt{\lambda}I \end{bmatrix} x \right\|_2^2.$$

Al definir

$$\widehat{y} := \begin{bmatrix} y \\ 0 \end{bmatrix} \quad \text{y} \quad \widehat{A} := \begin{bmatrix} A \\ \sqrt{\lambda}I \end{bmatrix}$$

podemos reformular el problema como

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|\widehat{y} - \widehat{A}x\|_2^2.$$

Solución usando mínimos cuadrados ordinarios

$$\underset{x\in\mathbb{R}^n}{\text{minimizar}} \quad \|\widehat{y} - \widehat{A}x\|_2^2.$$

La matriz $\widehat{A}$ es de $(m+n)\times n$ por lo que es **alta**: **siempre** tiene más filas que columnas

Además **siempre rango completo**: si existiese $z \in \mathbb{R}^n$ tal que $\widehat{A}z = 0$ entonces

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \widehat{A}z = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} z = \begin{bmatrix} Az \\ \sqrt{\lambda} z \end{bmatrix} \;\Rightarrow\; z = 0$$

Solución usando mínimos cuadrados ordinarios

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|\widehat{y} - \widehat{A}x\|_2^2.$$

Concluimos que se trata de un problema de **mínimos cuadrados ordinarios**

La **ecuación normal** es

$$\widehat{A}^\top \widehat{A}x = \widehat{A}^\top \widehat{y} \quad \Rightarrow \quad (A^\top A + \lambda I)x = A^\top y.$$

Podemos resolver el problema usando la **factorización QR** de $\widehat{A}$ si bien esto no es lo más eficiente en la práctica

UC | Chile

Solución usando la descomposición en valores singulares

Solución usando la descomposición en valores singulares

La **descomposición en valores singulares (SVD)** nos permite resolver el problema e interpretar su solución

Solución usando la descomposición en valores singulares

La **SVD completa** de una matriz $A \in \mathbb{R}^{m \times n}$ de rango r es la factorización $A = U\Sigma V^\top$ donde

- $U \in \mathbb{R}^{m \times m}$ con $U^\top U = UU^\top = I$
- $V \in \mathbb{R}^{n \times n}$ con $V^\top V = VV^\top = I$
- $\Sigma \in \mathbb{R}^{m \times n}$ de la forma

$$\Sigma = \begin{bmatrix} \Sigma_r & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix}$$

con $\Sigma_r = \mathbf{diag}(\sigma_1, \ldots, \sigma_r)$ y $\sigma_1 \geq \ldots \geq \sigma_r > 0$ son los **valores singulares**

Solución usando la descomposición en valores singulares

$$(A^\top A + \lambda I)x = A^\top y.$$

La SVD completa de A nos permite simplificar esta ecuación

Por una parte,

$$A^\top A = (U\Sigma V^\top)^\top (U\Sigma V^\top) = (V\Sigma^\top U^\top)(U\Sigma V^\top) = V\Sigma^\top \Sigma V^\top$$

de modo que, usando $VV^\top = I$, se tiene

$$A^\top A + \lambda I = V\Sigma^\top \Sigma V^\top + \lambda V V^\top = V(\Sigma^\top \Sigma + \lambda I)V^\top$$

Solución usando la descomposición en valores singulares

$$V(\Sigma^\top \Sigma + \lambda I)V^\top x = A^\top y$$

Por otra parte,

$$A^\top y = (U\Sigma V^\top)^\top y = V\Sigma^\top U^\top y$$

Solución usando la descomposición en valores singulares

$$V(\Sigma^\top \Sigma + \lambda I)V^\top x = V\Sigma^\top U^\top y$$

Ya que V es invertible, al definir la variable auxiliar $z = V^\top x$ el sistema se reduce a

$$(\Sigma^\top \Sigma + \lambda I)z = \Sigma^\top U^\top y$$

Solución usando la descomposición en valores singulares

$$(\Sigma^\top \Sigma + \lambda I) z = \Sigma^\top U^\top y$$

Este sistema es **diagonal**

$$\Sigma^\top \Sigma + \lambda I = \begin{bmatrix} \Sigma_r^\top \Sigma_r + \lambda I_r & 0_{r\times(n-r)} \\ I_{(n-r)\times r} & \lambda I_{n-r} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 + \lambda & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r^2 + \lambda & 0 & \dots & \vdots \\ 0 & \dots & 0 & \lambda & \dots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & \lambda \end{bmatrix}$$

Solución usando la descomposición en valores singulares

$$(\Sigma^\top \Sigma + \lambda I) z = \Sigma^\top U^\top y$$

Además

$$\Sigma^\top U^\top y = \begin{bmatrix} \sigma_1 u_1 \cdot y \\ \vdots \\ \sigma_r u_r \cdot y \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Solución usando la descomposición en valores singulares

$$(\Sigma^\top \Sigma + \lambda I) z = \Sigma^\top U^\top y$$

Por lo tanto, el sistema tiene la forma

$$\begin{bmatrix} (\sigma_1^2+\lambda) z_1 \\ \vdots \\ (\sigma_r^2+\lambda) z_r \\ \lambda z_{r+1} \\ \vdots \\ \lambda z_n \end{bmatrix} = \begin{bmatrix} \sigma_1 u_1 \cdot y \\ \vdots \\ \sigma_r u_r \cdot y \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \; z_i^\star = \begin{cases} \frac{\sigma_i}{\sigma_i^2+\lambda}(u_i \cdot y) & i \in \{1,\dots,r\} \\ 0 & i \in \{r+1,\dots,m\} \end{cases}$$

Solución usando la descomposición en valores singulares

Ya que $V^\top x^\star = z^\star$ se tiene

$$x^\star = \sum_{i=1}^{n} z_i^\star v_i = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda}(u_i \cdot y)v_i = \sum_{i=1}^{r} \frac{1}{\sigma_i}\frac{1}{1 + \lambda/\sigma_i^2}(u_i \cdot y)v_i$$

Esta representación muestra el efecto del valor λ

Cuando λ es **pequeño** el factor $1/\sigma_i$ predomina y genera **inestabilidad** cuando σ_i es muy pequeño

Cuando λ es **grande** el factor $1/\sigma_i$ es atenuado cuando σ_i es muy pequeño, **estabilizando** la solución

Solución usando la descomposición en valores singulares

Si definimos

$$U_r = \begin{bmatrix} u_1 & \dots & u_r \end{bmatrix} \quad \text{y} \quad V_r = \begin{bmatrix} v_1 & \dots & v_r \end{bmatrix}$$

entonces podemos representar matricialmente la solución como

$$x^\star = \sum_{i=1}^{r} \frac{1}{\sigma_i^2 + \lambda} (\sigma_i u_i \cdot y) v_i = V_r (\Sigma_r^2 + \lambda I_r)^{-1} \Sigma_r U_r^\top y$$

UC | Chile

La pseudoinversa

La pseudoinversa

¿Qué ocurre cuando $\lambda \to 0$? En tal caso,

$$x^\star = \sum_{i=1}^{r} \frac{1}{\sigma_i}(u_i \cdot y)v_i = V_r \Sigma_r^{-1} U_r^\top y.$$

Definimos la **pseudoinversa (de Moore-Penrose)** A^+ como

$$A^+ = V_r \Sigma_r^{-1} U_r.$$

Vemos que $A^+ \in \mathbb{R}^{n \times m}$

¿Qué propiedades tiene esta matriz?

La pseudoinversa

Propiedad	Representación matemática
Preserva información de columnas	$A^+AA^+ = A^+$
Inversa generalizada	$AA^+A = A$
Simetría de AA^+	$(AA^+)^\top = AA^+$
Simetría de A^+A	$(A^+A)^\top = A^+A$

La pseudoinversa

Cuando A es de **rango completo** podemos relacionar la pseudoinversa con

- la solución al problema de **mínimos cuadrados ordinarios** cuando A es **alta**
- la **solución de mínima norma Euclideana** cuando A es **ancha**

La pseudoinversa

Propiedad A rango completo	Representación matemática
Pseudoinversa para A alta	$A^+ = (A^\top A)^{-1} A^\top$
Pseudoinversa para A ancha	$A^+ = A^\top (A A^\top)^{-1}$
Solución de mínimos cuadrados ordinarios	$x^\star = A^+ y$
Solución de mínima norma Euclideana	$x^\star = A^+ y$

UC | Chile

Generalizaciones

Generalizaciones

En algunas aplicaciones se busca un balance entre **la magnitud del residuo** y **la magnitud de una transformación lineal del vector**

Si $L \in \mathbb{R}^{k \times n}$ buscamos encontrar el valor mínimo de

$$\|r(x)\|_2^2 + \lambda \|Lx\|_2^2$$

Generalizaciones

En este caso, podemos realizar la misma reducción previa al definir

$$\widehat{A} := \begin{bmatrix} A \\ \sqrt{\lambda} L \end{bmatrix}$$

La matriz $\widehat{A}$ es de $(m+k) \times n$ y será de **rango completo** cuando

$$\mathbf{nul}(L) \cap \mathbf{nul}(A) = \{0\}$$

Esto ocurre, por ejemplo, cuando $L \in \mathbb{R}^{n \times n}$ y es invertible

UC | Chile