Metodi Numerici per l'Informatica

Anthony

14 mar 2023

1 Regolarizzazione: introduzione

Abbiamo osservato problemi di fitting del tipo $y_i = ax_i + b$ che sono formalizzati come il seguente problema di minimizzazione:

$$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Abbiamo anche osservato che possiamo effettuare la regressione polinomiale in modo analogo. Ricordiamo che la regressione polinomiale è lineare nei parametri ma polinomiale rispetto ai dati:

$$y_i = b + \sum_{j=1}^{k} a_j x_i^j \quad \forall i = 1, \ldots, n$$

Usando la notazione matriciale, l'errore quadratico medio è scritto come segue:

$$\ell(\theta) = \|\mathbf{y} - \mathbf{X}\theta\|_2^2$$

settando il gradiente rispetto a θ pari a zero e risolvendo per θ otteniamo la seguente espressione:

$$\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

La quale è una soluzione in forma chiusa della regressione lineare. Osserviamo che θ non è esattamente uguale a zero, ma minimizza l'uguaglianza. In altre parole, θ è una soluzione approssimata che soddisfa la seguente espressione:

$$\mathbf{X}\theta \approx \mathbf{y}$$

In cui l'errore residuo $\|\mathbf{y} - \mathbf{X}\theta\|_2$ è il più piccolo possibile.

2 Equazioni normali

Consideriamo il seguente problema lineare:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Se esiste una soluzione lineare possiamo scrivere:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Ma se la soluzione lineare non esiste, dobbiamo risolvere un problema di approssimazione:

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

La cui possiamo riscrivere come segue:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

E la soluzione è quindi:

$$\mathbf{x} = \underbrace{(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T}_{\text{pseudo-inversa } \mathbf{A}^+}\mathbf{b}$$

La pseudo-inversa è anche chiamata l'*inversa di Moore-Penrose*. Essa viene utilizzata per risolvere problemi che prevedono un certo scarto quadratico medio.

3 Tipi di sistemi lineari

Esistono diversi tipi di sistemi lineari. Li identifichiamo con le seguenti categorie:

1. **Esatto**: n equazioni linearmente indipendenti e $m = n$ parametri. La matrice $\mathbf{A}$ è quindi quadrata.

   $$\text{Problema: } \mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{Soluzione: } \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

2. **Sovra-determinato**: n equazioni linearmente indipendenti e $m < n$ parametri. La matrice $\mathbf{A}$ è alta.

   $$\text{Problema: } \mathbf{A}\mathbf{x} \approx \mathbf{b} \quad \text{Soluzione: } \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

3. **Sotto-determinato**: n equazioni linearmente indipendenti e $m > n$ parametri. La matrice $\mathbf{A}$ è larga.

   $$\text{Problema: } \mathbf{A}\mathbf{x} \approx \mathbf{b} \quad \text{Soluzione: } \mathbf{???}$$

In particolare, quando un problema è sotto-determinato, esistono infinite soluzioni. Molte di queste, però, non sono valide. Ad esempio, sarebbero preferibili polinomi che seguono sinuosamente l'andamento dei nostri punti. La *regolarizzazione* è l'aggiunta di informazioni atte a risolvere problemi simili: un problema viene *regolarizzato* quando vengono aggiunte informazioni al problema e quindi ne vengono ridotte le sue soluzioni.

4 La regolarizzazione

La regolarizzazione è la chiave per risolvere problemi sotto-determinati aggiungendo più informazioni per restringere le possibili soluzioni del nostro problema. Idea: effettuiamo assunzioni generali e le scriviamo come termini dell'ottimizzazione. I regolarizzatori portano con loro diversi benefici:

- Impongono un certo comportamento della soluzione, come la sua sparsità o la sua smoothness;
- Riducono l'ammontare di dati necessari;
- Rendono i problemi di ottimizzazione più semplici da risolvere.

4.1 Regolarizzazione di Tikhonov

Supponiamo di voler minimizzare l'errore quadratico medio di un certo problema. Ad esso, aggiungiamo un altro termine. In particolare, in questo esempio, aggiungiamo una penalty L_2:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

Per qualche $\alpha > 0$. In questo esempio, stiamo cercando il vettore $\mathbf{x}$ che minimizzi il risultato dell'espressione a sinistra e dell'espressione a destra.
Oss: La $\mathbf{x}$ che comprende tutti zeri minimizza la penalty, ma non l'espressione a sinistra. Difatti, se $\alpha \to 0$ non stiamo affatto regolarizzando. Se $\alpha \to \infty$, invece, non stiamo tenendo conto dei dati, ovvero delle $\mathbf{x}$. Nel caso generale, dobbiamo scegliere un α, risolvere il problema e osservare la soluzione. In caso essa non ci piaccia, trovare un altro α.
La funzione ottenuta è convessa in $\mathbf{x}$; possiamo calcolarne il gradiente e porlo uguale a zero:

$$\nabla_{\mathbf{x}}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|x\|_2^2) = \mathbf{0}$$

Per linearità del gradiente otteniamo:

$$\nabla_{\mathbf{x}}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\nabla_{\mathbf{x}}\|x\|_2^2 = \mathbf{0}$$

$$= 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} + 2\alpha\mathbf{x} = 0$$

$$= \mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{A}^T\mathbf{b} + \alpha\mathbf{x} = 0$$

$$= \mathbf{A}^T\mathbf{A}\mathbf{x} + \alpha\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

$$= (\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I})\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

La soluzione è quindi scalata rispetto ad α ed è applicabile anche per problemi sovra-determinati. Per introdurre la regolarizzazione di Tikhonov tutto ciò che dobbiamo fare è aggiungere α per la diagonale di $\mathbf{A}^T\mathbf{A}$. Questo procedimento è anche chiamato *ridge regression*.

4.1.1 Esempio: deblurring

Supponiamo di voler ripristinare un'immagine sfocata nella sua versione originale.

(a) Sharp (b) Blurry

La nostra incognita è $\mathbf{x}$, l'immagine non sfocata, mentre conosciamo $\mathbf{x}_{\text{blurry}}$, l'immagine sfocata, e $\mathbf{G}$, ovvero la mappa lineare che ha sfocato l'immagine. Questo problema è risolvibile con la regolarizzazione di Tikhonov ai minimi quadrati, a patto che sappiamo quale operatore abbia sfocato la foto:

$$\min_{\mathbf{x}} \|\mathbf{x}_{\text{blurry}} - \mathbf{G}\mathbf{x}\|_2^2 + \alpha \|\mathbf{x}\|_2^2$$

4.2 Norme L_p

In generale, possiamo applicare diverse norme per la regolarizzazione, ad esempio la norma L_p:

$$\min_{x} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_p^p$$

p può essere qualsiasi numero:

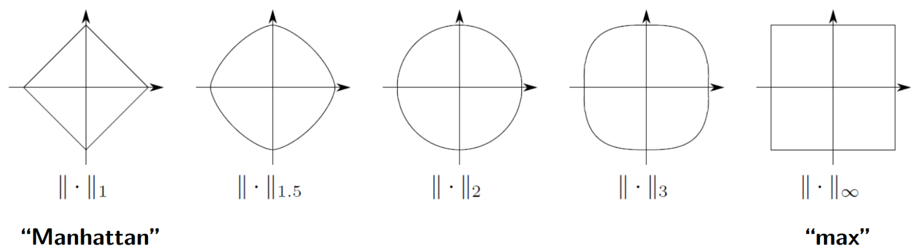
$$\|\mathbf{x} - \mathbf{y}\|_p = (|x_1 - y_1|^p + |x_2 - y_2|^p)^{\frac{1}{p}}$$

Generalizzando a $\mathbb{R}^k$:

$$\|\mathbf{x} - \mathbf{y}\|_p = (\sum_{i=1}^{k} |x_i - y_i|^p)^{\frac{1}{p}}$$

Questa definizione esprime il concetto di *distanza* L_p tra i vettori in $\mathbb{R}^k$.

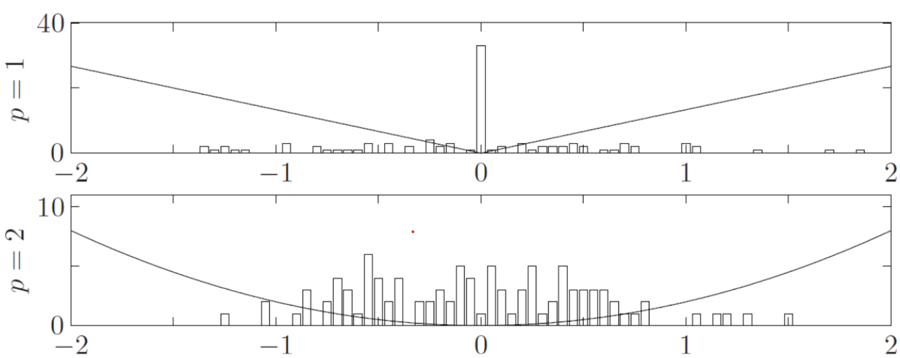
Circonferenza unitaria Diamo un'occhiata alla circonferenza unitaria applicando diverse norme L_p:



Ogni norma $\|\cdot\|_p$ esprime una *penalty* diversa.

4.3 Applicazione di Tikhonov con norme diverse

Tikhonov esprime un problema di minimizzazione, per cui dalla $\mathbf{x}$ dell'espressione ci aspettiamo che i valori siano piccoli e che i valori grandi siano scoraggiati. In particolare, l'applicazione di Tikhonov con la norma L_2 incoraggia ad avere valori tra 0 e ± 1. A seconda della decisione di p, i valori di $\mathbf{x}$ saranno penalizzati differentemente.



Possiamo osservare che la norma L_1 tende a preferire soluzioni sparse, ovvero sono presenti molti zeri nella soluzione, questo perché tutti i valori prendono una penalty pari al valore stesso.

4.4 Soluzioni sparse

Abbiamo osservato che la regolarizzazione con la norma L_1 è un'euristica per trovare soluzioni sparse. Ad esempio, consideriamo il seguente problema:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_1$$

Possiamo osservare che:

- Per $\alpha \approx 0$, questo problema è equivalente al problema dei minimi quadrati.

- Per $\alpha \gg 0$, la soluzione $\mathbf{x}$ conterrà molti zeri, e quindi sarà sparsa
- Se α è un valore meno estremo, stiamo facendo un compromesso tra la fedeltà dei dati e la sparsità di $\mathbf{x}$.

Tale problema non è differenziabile per la norma L_1. Non possiamo semplicemente computare il gradiente e settarlo a zero per trovare una soluzione, ma dobbiamo necessariamente approssimare il problema.

4.5 Approssimazione di funzioni non differenziabili

Proviamo ad applicare la norma L_1 a un vettore 1-dimensionale $\mathbf{x} = \begin{pmatrix} x \end{pmatrix}$ la sua norma L_1 sarà della forma:

$$f(\mathbf{x}) = |x|$$

Tale funzione non è differenziabile, ma possiamo riscriverla come segue:

$$f(\mathbf{x}) = \sqrt{|x|^2}$$

che ancora non è differenziabile. Però possiamo *ammorbidire* l'angolo del grafico aggiungendo ϵ molto piccolo:

$$f(\mathbf{x}) \approx \sqrt{|x|^2} + |\epsilon|$$

Questa espressione è un'approssimazione della norma L_1 di $\mathbf{x}$ ed è differenziabile. Ciò rende possibile la differenziazione di valori assoluti di cui non possiamo fare a meno.

4.6 Problemi sparsi

Finora abbiamo osservato soluzioni sparse a partire da un problema denso. In particolare, una matrice $\mathbf{A}$ è densa se la maggior parte dei numeri è diversa da zero. Se $\mathbf{A}$ non è densa, allora è sparsa. Consideriamo il seguente problema:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \alpha\rho(\mathbf{x}) \quad p \geq 1, \alpha \geq 0, \text{una funzione regolarizzatrice } \rho$$

Se la matrice $\mathbf{A}$ è sparsa, allora il problema si dice *problema sparso*. Ad esempio, la matrice $\mathbf{A}$ potrebbe essere una matrice tridiagonale:

$$\mathbf{A} = \begin{pmatrix} v_1 & w_1 & & & & \\ u_2 & v_2 & w_2 & & & \\ & u_3 & v_3 & w_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & u_{n-1} & v_{n-1} & w_{n-1} \\ & & & & u_n & v_n \end{pmatrix}$$

I grafi I grafi sono un altro esempio di problema sparso. Ricordiamo che un grafo può esser rappresentato attraverso una matrice di adiacenza: possiamo inserire un 1 in posizione i, j se il nodo i è connesso al nodo j, 0 altrimenti. Se il grafo non è particolarmente connesso, la matrice sarà sparsa.

Problemi sparsi: conclusione In generale, è bene avere un problema sparso poiché esistono algoritmi ad hoc molto efficienti.

5 Smoothing

Consideriamo il seguente problema: abbiamo un'immagine $\mathbf{x}$ e desideriamo sfocarla. Una soluzione a questo problema prevede il calcolo del problema di minimizzazione e come penalty viene aggiunta la norma del gradiente di $\mathbf{x}$ pesata da un certo α:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2 + \alpha \|\nabla \mathbf{x}\|_2^2$$

Intuitivamente, la norma L_2 promuove soluzioni *smooth*.

5.1 Smoothing quadratico

Esprimiamo termini della regolarizzazione come $\|\mathbf{Dx}\|$ in cui $\mathbf{D}$ è un qualche operatore di differenziazione. $\|\mathbf{Dx}\|$ rappresenta una misura della *variazione* o *smoothness* di $\mathbf{x}$:

$$\min_{\mathbf{x}} \underbrace{\|\mathbf{Ax} - \mathbf{b}\|_2^2}_{\text{data term}} + \alpha \underbrace{\|\mathbf{Dx}\|_2^2}_{\text{smoothness}}$$

Per esempio, assumiamo che $\mathbf{x} \in \mathbb{R}^n$ rappresenta una funzione in n punti. La sua derivata può esser approssimata come $\mathbf{\Delta x}$, in cui $\mathbf{\Delta}$ è la seguente mappa lineare:

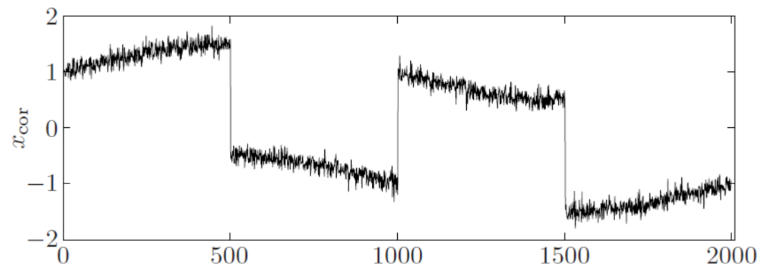
$$\mathbf{\Delta} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}$$

5.1.1 Denoising

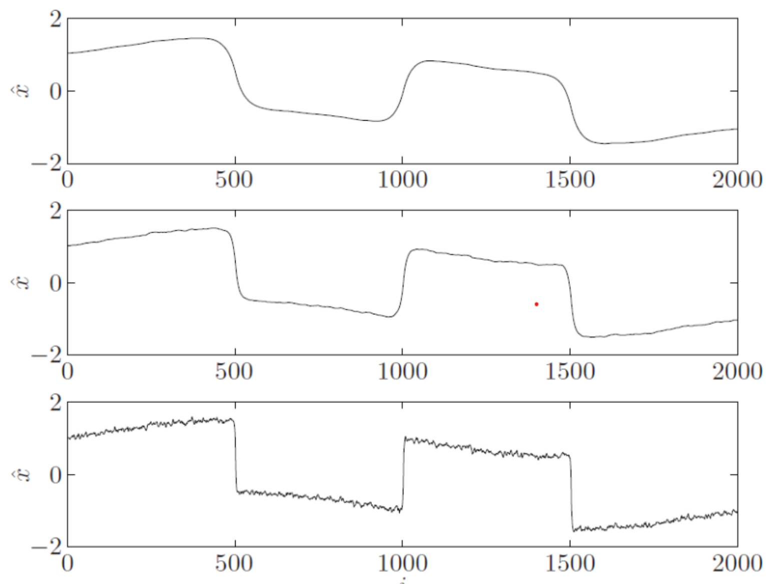
Un'applicazione è quella del denoising di un segnale audio. Supponiamo ci sia dato un segnale audio corrotto $\mathbf{x}_{\text{cor}}$ e vogliamo togliere il rumore dal segnale. Ottimizziamo quindi il seguente problema dello smoothing quadratico:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\text{cor}}\|_2^2 + \alpha \|\mathbf{\Delta x}\|_2^2$$

con $\mathbf{\Delta}$ definito come definito più sopra. Se il segnale originale è smooth, lo smoothing quadratico funziona bene tuttavia se consideriamo il seguente segnale rumoroso:



lo smoothing quadratico tratterà i vari salti di frequenza come rumore, attenuandoli:



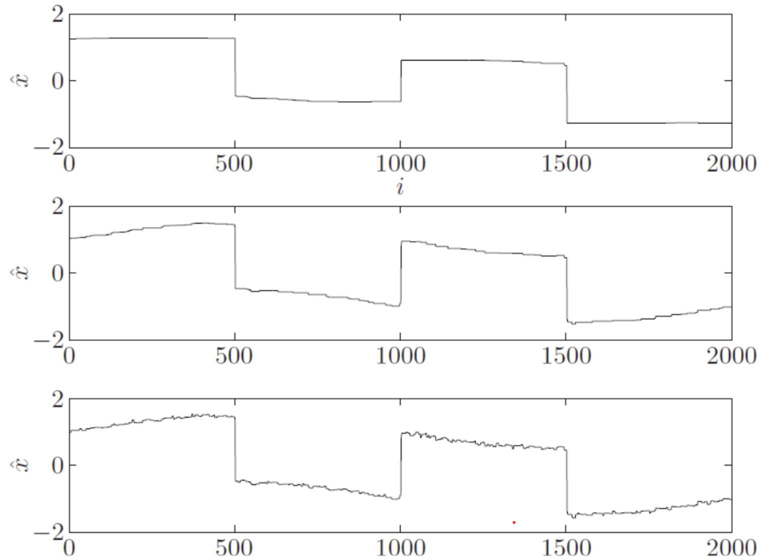
Per prevenire ciò, consideriamo la seguente funzione di smoothing:

$$\|\mathbf{\Delta x}\|_1 = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

E quindi il problema aggiornato:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\mathrm{cor}}\|_2^2 + \alpha \|\mathbf{\Delta x}\|_1$$

Come abbiamo già detto,utilizzare la norma L_1 corrisponde a trovare soluzioni sparse, per cui favorisce la sparsità del gradiente; qui non vogliamo affatto valori smooth. All'aumentare di α collasseremo in una retta.


2
0
−2
0
500
1000
1500
2000
i
$\hat{x}$