

Optimization for Machine Learning - Notes

Daniele Avolio - 242423

Academic Year 2023/2024

Contents

1	Introduzione	3
2	Programmazione non lineare	3
2.0.1	Caso di problemi senza vincoli	6
3	Approcci di classificazione	8
4	Separazione Lineare	9

List of Figures

1	Esempio di minimo locale stretto	4
2	Esempio di funzione convessa	5
3	Esempio di funzione non convessa	6

■ 1 Introduzione

Domanda 1.1. (Cosa significa costruire un classificatore?)

Significa costruire una superficie di separazione. Per farlo si allena un modello utilizzando dei dati etichettati, che prende il nome di **training set**. Le superfici di separazione ci aiutano a classificare nuovi dati non visti.

Esempio semplice: chi paga il mutuo e chi no.

Una superficie di separazione è definita come:

$$H(v, \gamma) = \{x \in \mathcal{R}^n | v^T x = \gamma\}$$

con:

- $v \in \mathcal{R}^n$ è un vettore, chiamato *normale*
- $\gamma \in \mathcal{R}$ è uno scalare, che è il *bias*

La funzione **sign** ci dice da che parte del piano si trova un punto. Cioè, dato un punto \bar{x} , se $\text{sign}(v^T \bar{x} - \gamma) \geq 0$ allora è un cliente che paga il mutuo, altrimenti no.

Domanda 1.2. (Dove interviene l'ottimizzazione quando si costruisce un classificatore? Perché serve?)

Il classificatore viene costruito andando a **minimizzare** una misura che indica quanto si sta sbagliando nel classificare i punti.

■ 2 Programmazione non lineare

Definition 2.1. (Minimo globale) Dato un punto $x^* \in \mathcal{R}^n$ si dice minimo globale se:

- $x^* \in X$, cioè il punto appartiene alla **regione ammissibile**
- $f(x^*) \leq f(x) \forall x \in X$, cioè per ogni punto della regione ammissibile, il valore di funzione obiettivo su x^* è minore uguale rispetto agli altri punti.

Notina: Definizione di programma lineare:

- $f(x) = c^T x$
- $X = \{x \in \mathcal{R}^n | Ax = b, x \geq 0\}$

Dove X è la regione ammissibile ed è un *poliedro*.

Definition 2.2. (Minimo locale)

Un punto $x^* \in X$ è un minimo locale per il problema P se:

- $x^* \in X$

- Esiste un vicinato N tale che $f(x^*) \leq f(x) \forall x \in X \cap N$. Cioé ogni punto della regione ammissibile intersecato col vicinato, e il valore x^* è sempre minore.

Il vicinato è un insieme di punti, non so come definito ma ok.

Definition 2.3. (Minimo locale stretto)

Un punto $x^* \in X$ è un minimo locale stretto per il problema P se:

- $x^* \in X$
- Esiste un vicinato N tale che $f(x^*) < f(x) \forall x \neq x^*, x \in X \cap N$.

Spiegazione al volo: Il minimo locale stretto è un minimo locale, ma non esistono altri punti che hanno lo stesso valore di funzione obiettivo.

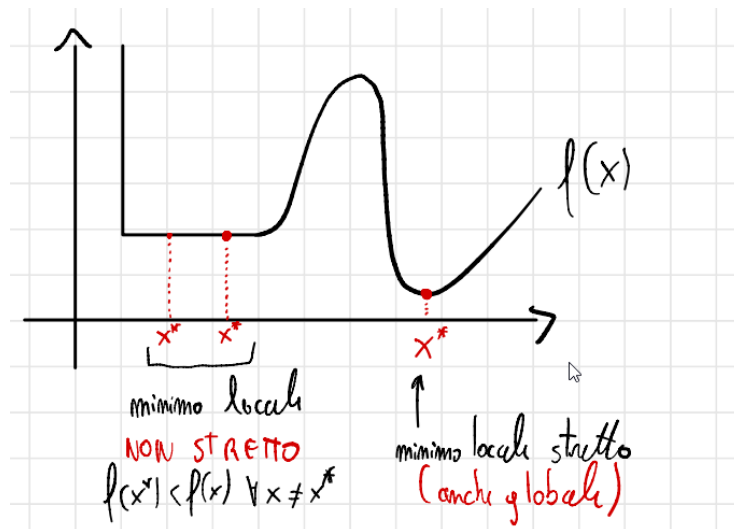


Figure 1: Esempio di minimo locale stretto

Nota: Se x^* è un minimo globale implica che x^* è un minimo locale.

Definition 2.4. (Combinazione convessa)

Dati $x^{(1)}$ e $x^{(2)}$ due punti $\in \mathbb{R}^n$, la combinazione convessa di $x^{(1)}$ e $x^{(2)}$ è un vettore:

$$\bar{x} = \lambda x^{(1)} + (1 - \lambda)x^{(2)}$$

con $\lambda \in [0, 1]$

Immagina una retta che unisce i due punti, con $\lambda = 0$ in $x^{(1)}$ e $\lambda = 1$ in $x^{(2)}$.

Definition 2.5. (Funzione convessa)

Data una funzione $f : \mathbb{R}^n \rightarrow \mathbb{R}$, f è **convessa** se per ogni coppia di punti $x^{(1)}, x^{(2)} \in \mathbb{R}^n$ e per ogni $\lambda \in [0, 1]$ vale che:

$$f(\lambda x^{(1)} + (1 - \lambda)x^{(2)}) \leq \lambda f(x^{(1)}) + (1 - \lambda)f(x^{(2)})$$

Cioè in italiano, il valore di funzione della combinazione dei due vettori è minore o uguale alla combinazione dei valori di funzione dei due vettori.

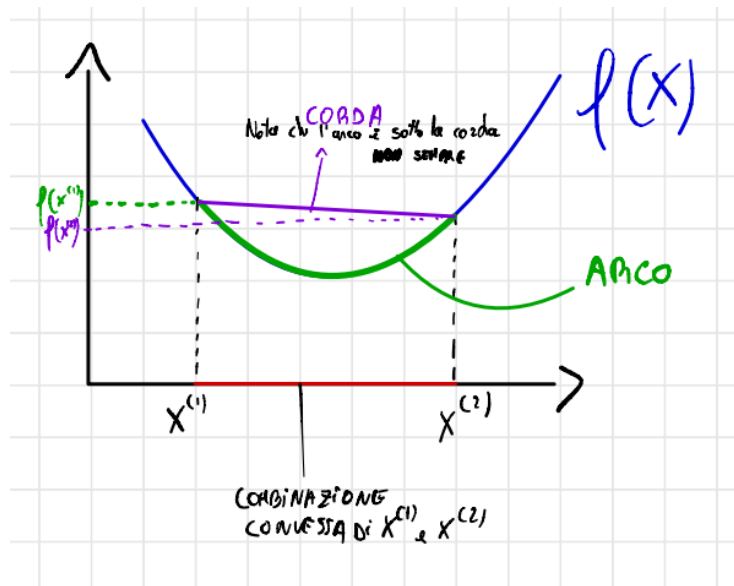


Figure 2: Esempio di funzione convessa

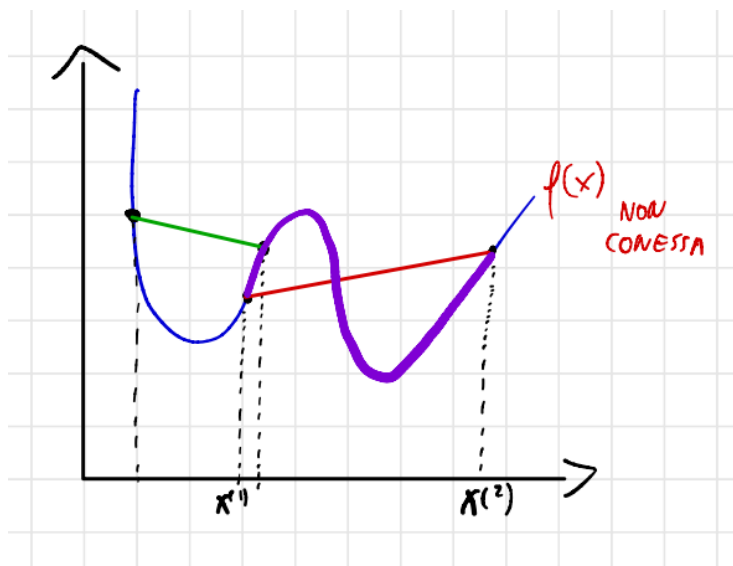


Figure 3: Esempio di funzione non convessa

Per capire, diciamo che la funzione è convessa se per ogni valore di funzione su un punto che è all'interno della combinazione convessa dei due punti, il valore di funzione è minore o uguale alla combinazione dei valori di funzione dei due punti.

Infatti, nel secondo esempio, ci sono dei punti tali per cui la funzione è maggiore (cioè sta sopra).

Domanda 2.1. (Quando un punto di un'insieme convesso è estremo?)

$\bar{x} \in X$ è un punto estremo di un'insieme convesso se NON ESISTE nessuna coppia di punti $x^{(1)}, x^{(2)} \in X$ e $\lambda \in (0, 1)$ tale che: $\bar{x} = \lambda x^{(1)} + (1 - \lambda)x^{(2)}$, per $\lambda \in]0, 1[$.

Banalmente, un punto è estremo se non è combinazione convessa di altri punti.

Nota: P è un programma convesso se f è una funzione convessa e X è un'insieme convesso. Questo ci serve saperlo perché in caso di **programma convesso** abbiamo che il minimo globale e locale **coincidono**.

Domanda 2.2. (Cosa cerchiamo con un problema di ottimizzazione?)

Cerchiamo il **minimo locale**, perché cercare il minimo globale fa parte di un'altra categoria di problemi, che sono quelli di **ottimizzazione globale**.

◆ 2.0.1 Caso di problemi senza vincoli

In questo caso, la regione ammissibile X coincide con \mathbb{R}^n .

$$P = \begin{cases} \min f(x) \\ f(x) : \mathbb{R}^n \rightarrow \mathbb{R} \end{cases}$$

Nota: Si fa un'assunzione. $f \in C^2$, cioè la funzione è due volte continuamente differenziabile. Quindi, C^2 è l'insieme di funzioni che ammettono prima e seconda derivate continue.

Questa assunzione ci permette di dire che $\bar{x} \in R^n \implies \nabla f(\bar{x})$ e $\nabla^2 f(\bar{x})$ esistono.

Vediamo come si applica il gradiente e la matrice hessiana.

Definition 2.6. (Gradiente)

Il gradiente di una funzione $f : R^n \rightarrow R$ è un vettore di dimensione n che contiene le derivate parziali della funzione rispetto alle sue variabili.

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Definition 2.7. (Matrice Hessiana)

La matrice hessiana di una funzione $f : R^n \rightarrow R$ è una matrice quadrata di dimensione n che contiene le derivate seconde parziali della funzione rispetto alle sue variabili.

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Esempio 2.1. $f(x) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$

Iniziamo dal gradiente.

$$\nabla f(x) = \begin{bmatrix} 8 + 2x_1 \\ 12 - 4x_2 \end{bmatrix}$$

Ora la matrice hessiana.

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

Dando un valore ad x , ad esempio $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, possiamo calcolare il gradiente e la matrice hessiana così:

$$\nabla f(x) = \begin{bmatrix} 10 \\ 8 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

■ 3 Approcci di classificazione

Abbiamo diversi approcci di classificazione.

- **Supervised learning:** Abbiamo un'insieme di dati che sono **etichettati**. Questo rappresenta il nostro **training set**. Il nostro obiettivo è fare predizioni sulle etichette di dati non ancora visti. Le etichette rappresentano la **classe**.
- **Unsupervised Learning:** I dati non hanno alcuna etichetta. Il nostro obiettivo è fare operazioni di **clustering**, ovvero raggruppare i dati in base a quanto sono simili tra loro.
- **Semisupervised Learning:** Abbiamo entrambi i tipi di dati (con e senza etichette). L'obiettivo è predire la label dei dati non etichettati.

Il modo in cui chiamo i dati all'interno del nostro dataset sono molteplici, tipo:

1. Datum
2. Object
3. Feature Vettore / Vettore delle caratteristiche
4. Punto

Definition 3.1. (*Classifier*)

Un classificatore è una **superficie di separazione** tra le classi.

■ 4 Separazione Lineare

Definition 4.1. (*Separazione Lineare*) Dati due insiemi $A = \{a_1, a_2, \dots, a_m\}$ e $B = \{b_1, b_2, \dots, b_k\}$. Due insiemi si dicono **linearmente separabili** \iff esiste un iperpiano $H(v, \gamma)$ che separa i due insiemi.

$$H(v, \gamma) = \{x \in \mathcal{R}^n | v^T x = \gamma\}$$

con:

- $v \in \mathcal{R}^n$ è un vettore
- $\gamma \in \mathcal{R}$ è uno scalare
- $v \neq 0$

Questo iperpiano, tale che:

$$v^T a_i \geq \gamma + 1 \wedge v^T b_j \leq \gamma - 1$$

per $i = 1, \dots, m$ e $j = 1, \dots, k$.

Nota e possibile domanda: Quando andiamo a classificare non teniamo conto del +1 e -1, perché vengono usati solo per costruzione. Quindi la disequazione conta solamente il valore di γ (nel lato destro).

Nota 2: I due insiemi A e B sono linearmente separabili \iff l'intersezione della loro copertura convessa è vuota.

$$\text{conv}(A) \cap \text{conv}(B) = \emptyset$$

Definition 4.2. (*Copertura Convessa*)

La copertura convessa di un'insieme X è l'insieme convesso più piccolo che lo contiene.

Un'insieme si dice convesso se per ogni coppia di punti $(x, y) \in X$ la combinazione di x e y è sempre all'interno dell'insieme X . Formalmente:

$$\forall x, y \in X, \forall \lambda \in [0, 1] \implies \lambda x + (1 - \lambda)y \in X$$

Implicazione ovvia, ma la copertura convessa di un'insieme convesso è l'insieme stesso. X convesso $\implies \text{conv}(X) = X$

Definition 4.3. (*Funzione Errore — Loss Function*)

Un punto $a_i \in A$ è **classificato correttamente** se

$$v^T a_i \geq \gamma + 1 \implies v^T a_i - \gamma - 1 \geq 0$$

Questo implica che a_i è **classificato erroneamente** se

$$v^T a_i - \gamma - 1 < 0 \implies -v^T a_i + \gamma + 1 > 0$$

L'errore di a_i è dato da:

$$\max\{0, -v^T a_i + \gamma + 1\} \geq 0$$

Analogamente, un punto $b_j \in B$ è **classificato correttamente** se

$$v^T b_j - \gamma + 1 \leq 0$$

. Questo implica che b_j è **classificato erroneamente** se

$$v^T b_j - \gamma + 1 > 0$$

L'errore di b_j è dato da:

$$\max\{0, v^T b_j - \gamma + 1\} \geq 0$$

References