

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
2.1	Condizioni di ottimalità	8
3	Eigen Values e Auto Vettori	8
4	Ottimizzazione senza vincoli	10
5	Ottimizzazione con vincoli	12
5.1	KKT Conditions	12
5.2	Teoremi delle KKT Conditions	13
6	Approcci di soluzione	15
7	Approcci di classificazione	19
8	Separazione Lineare	20
9	Separazione Poliedrale	21

List of Figures

1	Esempio di minimo locale stretto	4
2	Esempio di funzione convessa	5
3	Esempio di funzione non convessa	6
4	Funzione di errore nel caso poliedrico	23

■ 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.

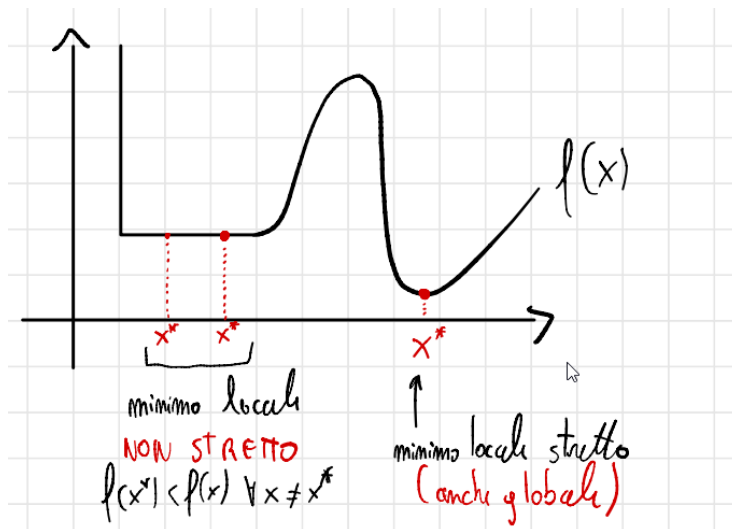


Figure 1: Esempio di minimo locale stretto

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.

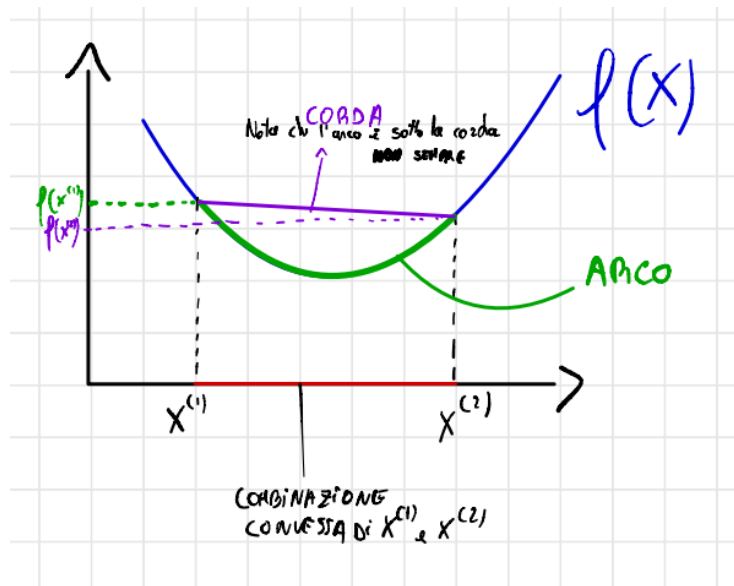


Figure 2: Esempio di funzione convessa

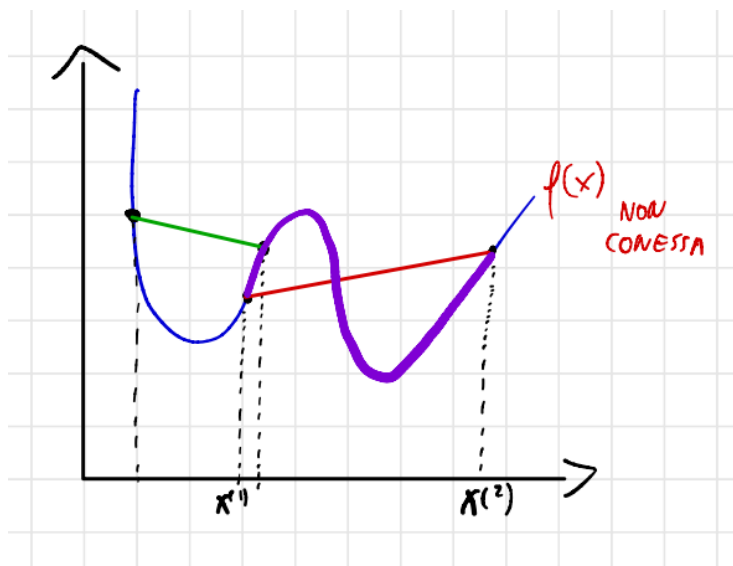


Figure 3: Esempio di funzione non convessa

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:

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

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

■ 2.1 Condizioni di ottimalità

Sono 3

Definition 2.8. (Prima condizione - condizione necessaria di primo ordine) x^* è un minimo locale implica che $\implies \nabla f(x^*) = 0$. Cioè, stiamo dicendo che x^* è un punto stazionario. Il fatto che il gradiente sia uguale a 0 è una condizione necessaria per fare in modo che x^* sia un minimo locale.

Nota: Se f è convessa, allora ogni punto stazionario è un **minimo globale**. (Pensa ad una funzione che è convessa ma ha più punti di minimo uguali)

Definition 2.9. (Seconda condizione - condizione necessaria di secondo ordine) x^* è un minimo locale stretto $\implies \nabla f(x^*) = 0$ e $\nabla^2 f(x^*)$ è positiva semidefinita. Lo definiamo dopo cosa Significa positiva semidefinita

Definition 2.10. (Terza condizione - condizione sufficiente di secondo ordine) Sia $x^* \in R^n$, sia $\nabla f(x^*) = 0$ e $\nabla^2 f(x^*)$ positiva definita $\implies x^*$ è un **minimo locale stretto**.

Definiamo cosa significa semidefinita ecc.

Sia A una matrice $R^{n \times n}$:

- **Positiva Semidefinita:** $\forall x \in R^n, x^T A x \geq 0$
- **Positiva Definita:** $\forall x \in R^n, x^T A x > 0$, con $x \neq 0$
- **Negativa Semidefinita:** $\forall x \in R^n, x^T A x \leq 0$
- **Negativa Definita:** $\forall x \in R^n, x^T A x < 0$, con $x \neq 0$

Negli altri casi, A è **indefinita**.

Nota: Non possiamo controllare queste definizioni perché R^n è infinito. Per questo motivo usiamo altre cose che si chiamano **eigen values**.

■ 3 Eigen Values e Auto Vettori

Sia A una matrice, λ uno scalare e x un vettore, con $x \neq 0$, diciamo che x è un **autovettore** e λ è un **autovalore**:

$$Ax = \lambda x$$

Semplicemente, matrice moltiplicato per vettore da lo stesso risultato per vettore moltiplicato per λ

Domanda 3.1. Come si calcolano gli Eigen Values?

Sapendo che $Ax = \lambda x$, possiamo riscriverlo come:

$$Ax - \lambda x = 0$$

Portiamo fuori x :

$$(A - \lambda I)x = 0$$

Dove I è la matrice identità.

Ora, per trovare gli eigen values, dobbiamo trovare i valori di λ tali che la matrice $(A - \lambda I)$ sia singolare. Cioè, il determinante deve essere uguale a 0.

$$\det(A - \lambda I) = 0$$

Fatto questo, si risolve l'equazione di secondo grado per trovare i valori di λ .

Esempio 3.1.

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\det(A - \lambda I) = 0$$

$$\det \begin{bmatrix} 2 - \lambda & 1 \\ 1 & 2 - \lambda \end{bmatrix} = 0$$

$$(2 - \lambda)^2 - 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = 0$$

$$\lambda_1 = 1 \wedge \lambda_2 = 3$$

Per calcolare il determinante di una matrice 2×2 si fa così:

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

Per calcolare il determinante di una matrice 3×3 si fa così:

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

Regola di Sarrus: Se si ha una matrice 3×3 , si può calcolare il determinante in un modo particolare.

Si ripete la matrice 3×3 in fila. Si calcolano queste cose:

- Somma dei prodotti delle prime 3 diagonali a partire da sinistra, verso destra
- Somma dei prodotti delle prime 3 diagonali a partire da destra, verso sinistra
- Sottraggo le due somme

Esempio 3.2.

$$\det \begin{bmatrix} 1 & 2 & 3 & 1 & 2 & 3 \\ 4 & 5 & 6 & 4 & 5 & 6 \\ 7 & 8 & 9 & 7 & 8 & 9 \end{bmatrix} = 1 \cdot 5 \cdot 9 + 2 \cdot 6 \cdot 7 + 3 \cdot 4 \cdot 8 - 3 \cdot 5 \cdot 7 - 2 \cdot 4 \cdot 9 - 1 \cdot 6 \cdot 8 = 0$$

Nota: Se la matrice A è una *matrice diagonale*, cioè una matrice in cui tutti gli elementi, tranne la diagonale, sono 0, allora gli *eigen values* sono gli elementi sulla diagonale.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\det(A - \lambda I) = 0$$

$$\det \begin{bmatrix} 1 - \lambda & 0 & 0 \\ 0 & 2 - \lambda & 0 \\ 0 & 0 & 3 - \lambda \end{bmatrix} = 0$$

$$(1 - \lambda)(2 - \lambda)(3 - \lambda) = 0$$

$$\lambda_1 = 1 \wedge \lambda_2 = 2 \wedge \lambda_3 = 3$$

Per quanto riguarda i segni delle matrici:

- A è una matrice **positiva semidefinita** \iff Eigen Values ≥ 0
- A è una matrice **positiva definita** \iff Eigen Values > 0
- A è una matrice **negativa semidefinita** \iff Eigen Values ≤ 0
- A è una matrice **negativa definita** \iff Eigen Values < 0

■ 4 Ottimizzazione senza vincoli

$$2x_1^3 - 3x_1^2 - 6x_1^2x_2 + 6x_1x_2^2 + 6x_1x_2$$

Calcoliamo il gradiente e la matrice hessiana

$$\nabla f(x) = \begin{bmatrix} 6x_1^2 - 6x_1 - 12x_1x_2 + 6x_2^2 + 6x_2 \\ -6x_1^2 + 6x_1^2 + 12x_1x_2 + 6x_1 \end{bmatrix}$$

Ora calcoliamo la matrice hessiana

$$\nabla^2 f(x) = \begin{bmatrix} 12x_1 - 6 - 12x_2 & -12x_1 + 12x_2 + 6 \\ -12x_1 + 12x_2 + 6 & 12x_1 \end{bmatrix}$$

Spieghiamo i passaggi per il calcolo della Hessiana

$$\frac{\partial^2 f}{\partial x_1^2} = 12x_1 - 6 - 12x_2$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = -12x_1 + 12x_2 + 6$$

Qui il $-12x_1 + 12x_2 + 6$ viene fuori dal seguente passaggio

In parole povere, prima si calcola la derivata parziale di f rispetto a x_1 , e poi si calcola la derivata di quel risultato rispetto a x_2 .

$$f(x_1, x_2) = 2x_1^3 - 3x_1^2x_2 + 6x_1x_2^2 + x_2^2 + 6x_1x_2$$

$$= 6x_1^2 - 6x_1x_2 + 6x_2^2 + 6x_2$$

$$= -12x_1 + 12x_2 + 6$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = -12x_1 + 12x_2 + 6$$

$$\frac{\partial^2 f}{\partial x_2^2} = 12x_1$$

Quindi, per spiegare come funziona.

Data una funzione, bisogna calcolare inizialmente il **gradiente**.

Applicando la **condizione necessaria di primo ordine** troviamo i punti stazionari, ovvero quelli in cui il $\nabla f(x) = 0$. Cioé, calcoli il gradiente e lo poni uguale a zero.

Ponendo il gradiente uguale a zero bisogna risolvere il sistema di equazioni per trovare i possibili punti stazionari. Una volta trovati, nel nostro esempio erano 4, bisogna esaminare i punti utilizzando la **condizione sufficiente di secondo ordine** e la **condizione necessaria di secondo ordine**.

Per farlo, si calcola l'Hessiana della funzione **in un punto**, per ogni punto. Ricorda bene come si calcola l'Hessiana, soprattutto quando compaiono due variabili.

Dopo aver calcolato l'hessiana e sostituito con il punto, bisogna **porre** il determinante della hessiana moltiplicata per l'identità con λ uguale a zero.

Calcolando il determinante e ponendolo uguale a zero si risolve l'equazione per trovare gli autovalori λ . In base al segno degli autovalori si può dire il "segno", della matrice. Controllando le condizioni necessarie e sufficienti di secondo ordine si può dire se il punto è un minimo locale, minimo locale stretto, massimo locale, massimo locale stretto, o punto sella.

■ 5 Ottimizzazione con vincoli

In Ottimizzazione con vincoli abbiamo due tipi di vincoli:

- Uguaglianza - $g(x) = 0$ - E
- Disuguaglianza - $g(x) \geq 0$ - I

Nota sui programmi quadratici: Se la funzione obiettivo è del tipo $f(x) = \frac{1}{2}x^T Mx + c^T x$, con M una matrice simmetrica che significa che $M = M^T$, e tutti i vincoli g_i sono funzioni lineari (sia di Uguaglianza che di Disuguaglianza), allora il problema è un **programma quadratico**.

Se invece la funzione è $f(x) = c^T x$, il problema quadratico diventa un programma lineare.

Definition 5.1. (Vincoli attivo)

Dato un punto $\bar{x} \in X$, un vincolo $g_i(\bar{x}) = 0$ si dice **vincolo attivo**.

Nota: Indichiamo $\mathcal{A}(\bar{x})$ l'insieme dei vincoli attivi in \bar{x} .

■ 5.1 KKT Conditions

Sono condizioni di ottimalità di primo ordine per i programmi con vincoli.

Vediamo in particolare cosa ci interessa:

Definition 5.2. (LICQ - Linear Independence Constraint Qualification)

Dato un punto $\bar{x} \in X$, LICQ regge in \bar{x} se l'insieme $\{\nabla g_i(\bar{x}), i \in \mathcal{A}(\bar{x})\}$ cioè l'insieme dei vincoli attivi per \bar{x} , deve essere costituito solamente da vettori linearmente indipendenti.

Definition 5.3. (Funzione Lagrangiana)

Dato un vettore $\lambda \in \mathbb{R}^{|E|+|I|}$ chiamato vettore dei moltiplicatori Lagrangiani, diciamo che la funzione lagrangiana di P è:

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in E} \lambda_i g_i(x) - \sum_{i \in I} \lambda_i g_i(x)$$

con $\lambda \geq 0 \forall i \in I$.

Se vogliamo fare un esempio, ecco la spiegazione di come si lavora.

Data una regione ammissibile, quindi un insieme di vincoli, analizziamo prendendo un punto \bar{x} come si comportano i vincoli.

Controlliamo quali sono i vincoli che si attivano, ovvero quando la funzione $g_i(\bar{x}) = 0$.

Prendiamo questi vincoli e calcoliamo il gradiente del vincolo, ovvero $\nabla g_i(\bar{x})$. Se abbiamo ancora delle variabili dopo aver fatto il gradiente, sostituiamo alla x che compare nel gradiente il punto \bar{x} .

Poi, dopo aver calcolato questi valori, inseriamo tutti i gradienti in una matrice, chiamata B . Bisogna controllare che i gradienti siano linearmente indipendenti, e

per comodità possiamo calcolare il **determinante** della matrice e controllare che sia $\neq 0$.

Se ad occhio si vede che dei gradienti sono linearmente dipendenti, allora si può dire direttamente che *LICQ* non reggono.

■ 5.2 Teoremi delle KKT Conditions

Ci sono delle condizioni da rispettare:

Sia x^* un minimo locale per il problema P e che le *LICQ* reggono. Allora possiamo dire che $\exists \lambda^*$ tale che:

$$KKT - Conditions = \begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) &= 0 \\ g_i(x^*) &= 0 \quad \forall i \in E \\ g_i(x^*) &\geq 0 \quad \forall i \in I \\ \lambda_i^* &\geq 0 \quad \forall i \in I \\ \lambda_i^* g_i(x^*) &= 0 \quad \forall i \in E \cup I \end{cases}$$

Una nota, che non sappiamo a cosa serve ma è importante.

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in E} \lambda_i g_i(x) - \sum_{i \in I} \lambda_i g_i(x)$$

Questa formula ci dice che la funzione lagrangiana è la funzione obiettivo meno la sommatoria dei vincoli moltiplicati per i moltiplicatori lagrangiani.

Quando applichiamo il gradiente rispetto ad x :

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in E} \lambda_i^* \nabla g_i(x^*) - \sum_{i \in I} \lambda_i^* \nabla g_i(x^*)$$

Ora, **nota** importante: Se $g_i(x^*) = 0$, allora serve che $\lambda^* = 0$, questo torna utile per l'ultima condizione di prima, ovvero $\lambda_i^* g_i(x^*) = 0 \quad \forall i \in E \cup I$. Perché questo implica che:

$$\begin{aligned} \implies \nabla_x \mathcal{L}(x^*, \lambda^*) &= \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla g_i(x^*) = 0 \\ \implies \nabla f(x^*) &= \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla g_i(x^*) \end{aligned} \tag{1}$$

In questo modo possiamo trovare il valore dei λ_i^* .

Solitamente, quando si lavora con un esempio, possiamo avere *diversi punti*. Si parte **verificando le LICQ** e, successivamente, si verificano le **KKT Conditions**. Partiamo dalle *LICQ* perché le *KKT conditions* hanno bisogno di avere le *LICQ* che reggono per quel punto per trovare il λ^* .

Per le **LICQ**:

Dato un punto controlliamo quali sono i vincoli $g_i(x) \in \mathcal{A}(x)$ che si attivano, ovvero quando la funzione $g_i(x) = 0$. Trovato questo insieme si calcola il

gradiente per ogni vincolo attivo. Successivamente si controlla che i vincoli siano linearmente indipendenti tra loro e, se lo sono, allora le LICQ reggono.

Per le **KKT Conditions**: Si calcola inizialmente la funzione lagrangiana. Questo è dato dalla formula che abbiamo visto prima, con la funzione obiettivo meno la sommatoria dei vincoli moltiplicati per i moltiplicatori lagrangiani.

Successivamente calcoliamo il gradiente della funzione lagrangiana rispetto ad x e lo poniamo uguale a zero. Ci ritroveremo ad avere un sistema di equazioni con λ_i come incognita. Per trovare il valore di λ_i bisogna risolvere il sistema di equazioni.

Dopo aver trovato il valore di λ^* si controlla che ogni valore di λ_i^* sia maggiore o uguale a zero, solamente per i vincoli di disuguaglianza.

Nota importante per tanti λ : Se in una regione ammissibile abbiamo tanti vincoli, il λ^* da trovare avrà tanti valori quanti i vincoli. Abbiamo però modo di semplificare il calcolo di questi. Per la proprietà della **complementarietà** abbiamo che:

$$\lambda_i * g_i(x^*) = 0 \quad \forall i \in E \cup I$$

. Ora, se sappiamo già che $g_i(x^*) > 0$, allora λ_i^* deve **PER FORZA** essere uguale a zero.

Ad esempio, se su 5 vincoli solo 2 sono attivi, al calcolo della funzione lagrangiana avremo 5 moltiplicatori lagrangiani, ma solamente 2 saranno diversi da zero perché i restanti 3 vincoli non sono attivi. Possiamo rimuovere dall'equazione per semplificare i calcoli.

Implicazione diretta: Se le LICQ reggono:

- x^* è un minimo locale \implies le KKT Reggono
- Se le KKT non reggono $\implies x^*$ non è un minimo locale

Notiamo che le KKT sono **condizioni necessarie** ma non sufficienti.

■ 6 Approcci di soluzione

Quando partiamo da un punto iniziale x^0 possiamo avere più casi di problemi.

- Caso senza vincoli $\rightarrow \in \mathbb{R}^n$
- Caso con vincoli. Questo caso si divide in due:
 - x potrebbe non appartenere alla regione ammissibile $x \notin X$
 - x appartiene alla regione ammissibile $x \in X$

Domanda 6.1. *Quando cerchiamo un nuovo punto, quando ci fermiamo?
Questo dipende dal tipo di problema che stiamo risolvendo.*

- *Caso senza vincoli: Ci fermiamo quando troviamo un punto stazionario, ovvero quando dato un punto x , il gradiente della funzione nel punto è uguale a 0 $\nabla f(x) = 0$.*
- *Caso con vincoli: Ci fermiamo quando troviamo un punto che soddisfa le KKT Conditions.*

Abbiamo due metodi di soluzione: **Metodi di ricerca lineari** che possono essere a loro volta **esatti** e **inesatti**, oppure il **Metodo della regione di confidenza**.

Definition 6.1. *Metodi di ricerca lineari*

Un nuovo punto x^{k+1} viene calcolato usando $x^k + \alpha_k d^k$, con α_k un passo di ricerca e d^k una direzione di ricerca.

Nota: $\alpha \in \mathbb{R} > 0$ e $d \in \mathbb{R}^n$.

Ad esempio, il simplesso è un metodo di ricerca lineare.

Domanda 6.2. *Qual è la differenza tra esatto e inesatto?*

Il calcolo del nuovo punto x^{k+1} può essere esatto o inesatto. Se è esatto, è risolto esattamente, se è inesatto è calcolato in modo approssimato.

Definition 6.2. *Regione di confidenza*

In questo metodo un nuovo punto x^{k+1} è calcolato usando il vecchio punto e una direzione.

La direzione d^k è calcolata risolvendo un sotto problema del tipo:

$$P_{TR} = \begin{cases} \min_d m_k(x^k + d^k) \\ x^k + d \in X \end{cases}$$

Dove si indica con m_k la funzione modello.

Quando si approssima per d , diciamo che è approssimata bene se la differenza tra la funzione originale e la funzione modello, nel punto nuovo calcolato è bassa. Altrimenti, è una cattiva approssimazione.

Definition 6.3. *Norma di un vettore*

Si indica con *norma del vettore* un mapping tra $R^n \rightarrow R^+$ che soddisfa le seguenti proprietà:

- $\|x\| = 0 \iff \vec{x} = 0$
- $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \quad \forall x, y \in R^n$
- $\|\alpha \vec{x}\| = |\alpha| \|\vec{x}\| \quad \forall \alpha \in R \wedge x \in R^n$

Ci sono varie norme, ma a noi interessa la norma L_2

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

Nota: Se $x \in R$, allora $\|x\| = |x|$.

Nota 2: La norma non è differenziabile perché la radice quadrata non è differenziabile in 0.

Nota 3: Possiamo scrivere $\|x\|_2^2 = x^T x$.

Caso senza vincoli

$$P = \begin{cases} \min_x f(x) \\ f : R^n \rightarrow R \\ f \in C^2 \end{cases}$$

Definition 6.4. *(Direzione di discesa)*

Dato un $\bar{x} \in R^n$, il vettore \bar{d} è una *direzione di discesa* di f al punto \bar{x} se esiste un $\bar{\alpha} > 0$ tale che:

$$f(\bar{x} + \alpha \bar{d}) < f(\bar{x}) \quad \forall \alpha \in (0, \bar{\alpha})$$

Nota: Se questa \bar{d} esiste e $\nabla f(\bar{x})^T \bar{d} < 0$, allora la direzione è una direzione di discesa. Se siamo nel caso in cui f è convessa, allora l'inverso è vero.

$$\nabla f(\bar{x})^T \bar{d} < 0 \iff \bar{d} \text{ è una direzione di discesa}$$

Definition 6.5. *(Direzione di steepest descent)*

Dal teorema di Taylor sappiamo che possiamo approssimare $f(x^k + d)$ con $f(x^k) + \nabla f(x^k)^T d$. Vogliamo minimizzare questa approssimazione, per d .

$$P_K = \begin{cases} \min_d f(x^k) + \nabla f(x^k)^T d \\ \frac{1}{2} \|d\|_2^2 = \frac{1}{2} \end{cases}$$

Ma la prima parte possiamo semplificare perché non compare la d .

Il nostro obiettivo è trovare d . Come facciamo? Possiamo usare il metodo dei moltiplicatori lagrangiani.

$$\mathcal{L}(d, \lambda) = \nabla f(x^k)^T d - \lambda \left(\frac{1}{2} \|d\|_2^2 - \frac{1}{2} \right)$$

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla f(x^k) - \lambda d = 0$$

$$\nabla f(x^k) = \lambda d$$

Si assume che $\lambda \neq 0$ perché se fosse zero, allora $\nabla f(x^k) = 0$ e quindi x^k sarebbe un punto stazionario.

$$d = \frac{\nabla f(x^k)}{\lambda}$$

Ora, dal vincolo sappiamo che $\|d\|^2 = 1$, quindi:

$$\frac{\|\nabla f(x^k)\|^2}{\lambda^2} = 1$$

Da qui troviamo i moltiplicatori lagrangiani.

$$\lambda^2 = \|\nabla f(x^k)\|^2 \rightarrow \lambda = \pm \|\nabla f(x^k)\|$$

Siccome siamo in un contesto di minimizzazione, prendiamo il valore negativo.

$$d^* = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

Nota: $\nabla f(x^k)^T d^* = \frac{-\nabla f(x^k)^T \nabla f(x^k)}{\|\nabla f(x^k)\|} = -\|\nabla f(x^k)\| < 0$.

Quando hai un esempio svolto, per capire se effettivamente la **step size** cioè α è buono, controlla se effettivamente il nuovo punto ha un valore inferiore.

$$f(x^k + \alpha d^k) < f(x^k)$$

Definition 6.6. Metodo di Newton

Il problema è sempre del tipo:

$$P = \begin{cases} \min_x f(x) \\ f : \mathbb{R}^n \rightarrow \mathbb{R} \\ f \in C^2 \end{cases}$$

Sempre dal teorema di Taylor, possiamo approssimare:

$$f(x^k + d) \approx f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T \nabla^2 f(x^k) d$$

Assumiamo che $\nabla^2 f(x^k)$ sia definita positiva.

Il problema che vogliamo risolvere è:

$$P_K = \left\{ \min_d m_k(d) \right.$$

Risolviamo il problema:

$$\nabla_d m_k(d) = \nabla f(x^k) + \nabla^2 f(x^k)d = 0$$

$$\nabla f(x^k) = -\nabla^2 f(x^k)d$$

che è positiva definita

$$d = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

che è la direzione di Newton

$$\nabla f(x^k)^T d = \nabla f(x^k)^T [-\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$$

$$= -\nabla f(x^k)^T \nabla^2 f(x^k)^{-1} \nabla f(x^k) < 0$$

è una direzione di discesa

Nota: La parte interna senza il $-$ deve essere > 0

Note 2:

- $m_k(0) = f(x^k)$
- $\nabla m_k(0) = \nabla f(x^k)$
- $\nabla^2 m_k(0) = \nabla^2 f(x^k)$

■ 7 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 7.1. (*Classifier*)

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

■ 8 Separazione Lineare

Definition 8.1. (*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, la normale del piano
- $\gamma \in \mathcal{R}$ è uno scalare, il bias
- $v \neq 0$

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 8.2. (*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 8.3. (*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:

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

La funzione errore totale che ci viene da: $\min_{v, \gamma} f(v, \gamma)$ è:

$$\min_{v, \gamma} f(v, \gamma) = \frac{1}{m} \sum_{i=1}^m \max\{0, -v^T a_i + \gamma + 1\} + \frac{1}{k} \sum_{j=1}^k \max\{0, v^T b_j - \gamma + 1\}$$

Cioè, la media degli errori di classificazione dei punti di A e B .

Questa funzione è **non smooth e convessa**. Possiamo trasformare questo in un programma lineare.

Consideriamo:

- $\xi_i = \max\{0, -v^T a_i + \gamma + 1\}$ per $i = 1, \dots, m$
- $\psi_j = \max\{0, v^T b_j - \gamma + 1\}$ per $j = 1, \dots, k$

Possiamo riscrivere il problema nel seguente modo:

$$LP = \begin{cases} \min_{v, \gamma, \xi, \psi} \frac{1}{m} \sum_{i=1}^m \xi_i + \frac{1}{k} \sum_{j=1}^k \psi_j \\ \xi_i \geq -v^T a_i + \gamma + 1 \quad \forall i = 1, \dots, m \\ \psi_j \geq v^T b_j - \gamma + 1 \quad \forall j = 1, \dots, k \\ \xi_i, \psi_i \geq 0 \quad \forall i = 1 \end{cases}$$

■ 9 Separazione Poliedrale

La separazione poliedrale è un'estensione della separazione lineare. In questo caso consideriamo un problema che ha almeno $h > 1$ iperpiani di separazione $H(v^j, \gamma_j)$ tale che:

1. Per ogni punto positivo $a_i \in A$ e ogni iperpiano $j = 1, \dots, h$ vale che:

$$v^j T a_i \geq \gamma_j - 1$$

Cioè il punto è "dentro" l'iperpiano.

-
2. Per ogni punto negativo $b_i \in B$ esiste almeno 1 iperpiano $j \in \{1, \dots, h\}$ tale che:

$$v^j T b_i \geq \gamma_j + 1$$

Cioé il punto è "fuori" almeno 1 iperpiano.

Nota: A è h-pliedricamente separabile da $B \iff \text{conv}(A) \cap B = \emptyset$.

Nota 2: Se $h = 1 \rightarrow$ problema lineare.

Definition 9.1. *Funzione errore nel caso poliedrico:*

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

$$v^j T a_i \leq \gamma_j - 1 \quad \forall j = 1, \dots, h$$

$$v^j T a_i - \gamma_j + 1 \leq 0 \quad \forall j = 1, \dots, h$$

$$\max_{j=1, \dots, h} \{v^j T a_i - \gamma_j + 1\} \leq 0 \quad \forall j = 1, \dots, h$$

Ed è classificato erroneamente se:

$$\max_{j=1, \dots, h} \{v^j T a_i - \gamma_j + 1\} > 0 \quad \forall j = 1, \dots, h$$

In totale, l'errore di a_i è dato da:

$$\max_{j=1, \dots, h} \{0, v^j T a_i - \gamma_j + 1\}$$

Per quanto riguarda i punti negativi $b_j \in B$, viene classificato correttamente se $\exists j \in \{1, \dots, h\}$ tale che:

$$v^j T b_i \geq \gamma_j + 1 \rightarrow v^j T b_i - \gamma_j - 1 \geq 0 \rightarrow -v^j T b_i + \gamma_j + 1 \leq 0$$

Analogamente, viene classificato erroneamente se:

$$-v^j T b_i + \gamma_j + 1 > 0 \quad \forall j = 1, \dots, h$$

Cioè, se $\min_{j=1, \dots, h} \{-v^j T b_i + \gamma_j + 1\} > 0$.

L'errore di b_i è dato da:

$$\max_{j=1, \dots, h} \{0, -v^j T b_i + \gamma_j + 1\}$$

*Notiamo che questa funzione è **più difficile da minimizzare**.*

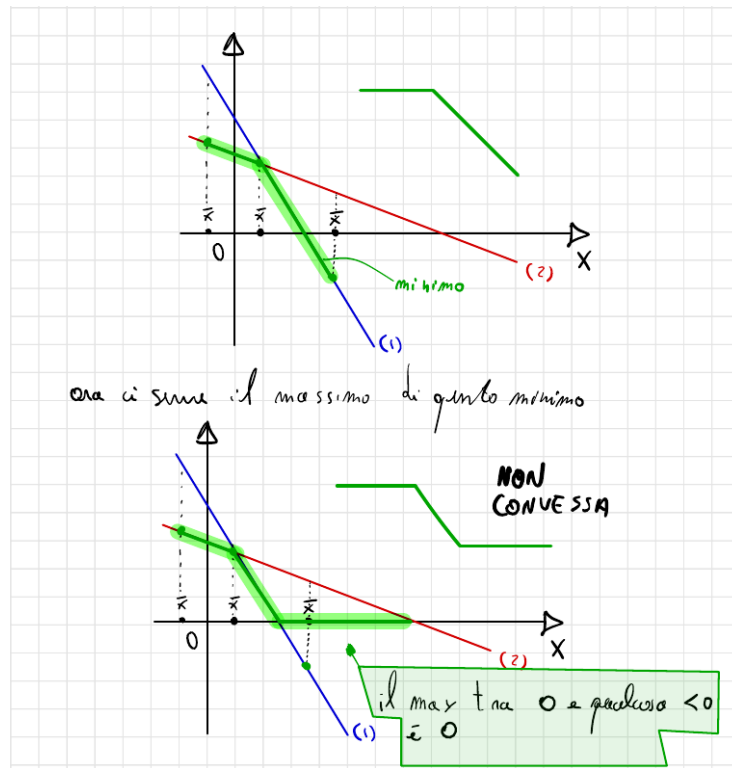


Figure 4: Funzione di errore nel caso poliedrico

La funzione di errore totale è data da:

$$f(v^1, v^2, \dots, v^i, \gamma_1, \gamma_2, \dots, \gamma_h) = \frac{1}{m} \sum_{i=1}^m \max_{j=1, \dots, h} \{0, v^{jT} a_i - \gamma_j + 1\} + \frac{1}{k} \sum_{j=1}^k \max_{i=1, \dots, h} \{0, -v^{jT} b_i + \gamma_j + 1\}$$

Questa funzione è **non convessa**.

References