Indice

PICCIALLI

1 Classificazione supervisionata

Nella **classificazione supervisionata** sono noti a priori, tra i dati di training e di set i pattern che rappresentano le diverse classi e si vuole determinare un modello matematica che definisca la classe di appartenenza.

2 Classificazione non supervisionata

Nella **classificazione non supervisionata** non sono noti a priori i pattern rappresentativi delle classi. Si vuole determinare il numero di classi di similitudine e un modello matematico che dato un pattern dello spazio delle caratteristiche, definisca la corrispondente classe di appartenenza.

3 Regressione

Nella **regressione** sono note coppie di pattern/target rappresentative di una funzione incognita a valori reali. Si vuole determinare una funzione analitica che approssima la funzione incognita.

4 Problemi di ottimizzazione continua definisi su $\mathbb{R}^n$

Nei problemi continui a dimensione finita, lo spazio delle variabile o **spazio delle feature**, è uno spazio lineare in $\mathbb{R}^n$. In particolare si ha un problema del tipo:

$$\min f(x)$$

$$x \in S \subseteq \mathbb{R}^n$$

$$f\colon \mathbb{R}^n \to \mathbb{R} := \textbf{funzione obiettivo}$$

$$S := \textbf{insieme ammissibile}$$

Questi problemi possono essere classificate in base alla struttura dell'insieme ammissibile e alle ipotesi sulla funzione obiettivo:

- Se $S = \mathbb{R}^n$ si parla di **ottimizzazione non vincolata**;
- Se $S \subset \mathbb{R}^n$ si parla di **ottimizzazione vincolata**. Il problema più comune di questa categoria è quello in cui S è descritto attraverso un **insieme finito di vincoli di uguaglianza e diseguaglianza**:

  $$S = \{x \in \mathbb{R}^n : g(x) \leqslant 0, h(x) = 0\}$$

  In cui: $g: \mathbb{R}^n \to \mathbb{R}^m \quad e \quad h: \mathbb{R}^n \to \mathbb{R}^p$ vettori di funzioni assegnate.

5 Condizioni di Minimo

5.1 Punto di minimo globale

Un punto $x^* \in S$ è detto punto di **minimo globale** di f su S se:

$$f(x^*) \leqslant f(x) \quad \forall x \in S$$

e il valore $f(x^*)$ è il minimo globale di f su S.

Si dice che $x^* \in S$ è un punto di **minimo globale stretto** di f su S se:

$$f(x^*) < f(x) \quad \forall x \in S, x \neq x^*$$

5.2 Punto di minimo locale

Un punto $x^* \in S$ è detto di **minimo locale** di f su S se esiste un intorno $B(x^*, \rho)$ con $\rho > 0$ tale che:

$$f(x^*) \leqslant f(x) \quad \forall x \in S \cap B(x^*, \rho)$$

Un punto $x^* \in S$ è detto di **minimo locale stretto** se esiste un intorno $B(x^*, \rho)$ con $\rho > 0$ tale che:

$$f(x^*) < f(x) \quad \forall x \in S \cap B(x^*, \rho)$$

5.3 Punto di minimo locale isolato

Dato un punto x^* è un punto di **minimo locale isolato** quando è un punto di minimo locale, ma l'intorno non è totalmente contenuto in S.

5.4 Convessità

Funzione convessa

Un funzione $f(x)$ è **convessa** nell'insieme S se:

$$\forall x, y \in S \quad \forall \lambda \in [0,1] \to f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y)$$

In cui:

- $f(\lambda x + (1-\lambda)y) :=$ segmento che connette x,y;

- $\lambda f(x) + (1-\lambda) f(y) :=$ segmento che connette $f(x)$ e $f(y)$

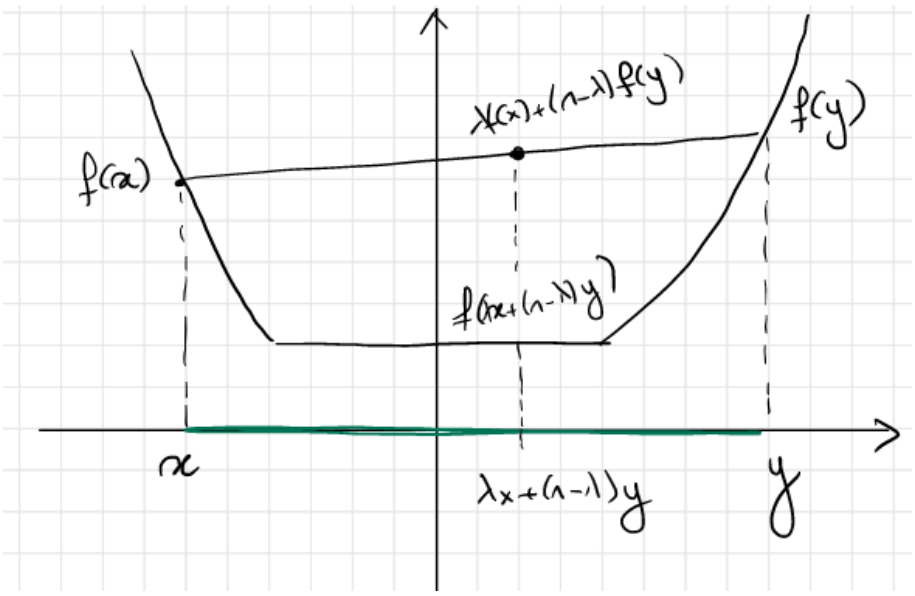


Figura 1.

Inoltre, $f(x)$ è **strettamente convessa** su S se $\forall x, y \in S, \forall \lambda \in [0,1]$:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

Oss.:

$$\text{Convessità} \nRightarrow \text{Esistenza di minimo}$$

es. f(x)=e^x, è strettamente convessa su $\mathbb{R}$, ma non ammette minimo globale.

Insieme convessi

S è un **insieme convesso** se $\forall x, y \in S \rightarrow \lambda x + (1-\lambda)y \in S \quad \forall \lambda \in [0,1]$

Corollario

Se $g(x)$ è convesso $\rightarrow S$: $\{x \in \mathbb{R}: g(x) \leqslant 0\}$ è convesso

Dimostrazione

Occorre dimostrare che $\forall x, y \in S \quad \lambda x + (1-\lambda)y \in S \quad \forall \lambda \in [0,1]$

Quindi:

$$x, y \in S \rightarrow g(x) \leqslant 0, g(y) \leqslant 0 \rightarrow \lambda x + (1-\lambda)y \in S \wedge g(\lambda x + (1-\lambda)y \leqslant 0$$

Poiché g è convessa$\rightarrow g(\lambda x + (1-\lambda)y) \leqslant \lambda g(x) + (1-\lambda) g(y)$:

- $(1-\lambda) \geqslant 0 \quad$ poiché $\lambda \in [0,1]$
- $g(y), g(x) \leqslant 0 \quad$ per ipotesi;

Quindi:

$$g(\lambda x + (1-\lambda)y) \leqslant \lambda g(x) + (1-\lambda) g(y) \leqslant 0 \rightarrow \lambda x + (1-\lambda)y \in S \rightarrow \textbf{Insieme convesso}$$

Corollario

Se $S_1, \ldots, S_n$ sono convessi $\rightarrow S_1 \cap \cdots \cap S_n$ è convesso.

Dimostrazione

$x, y \in S_1 \Longleftrightarrow x, y \in S$

Se S_i convesso $\rightarrow \lambda x + (1-\lambda)y \in S_i \quad \forall \lambda \in [0,1] \, \forall i \longrightarrow \lambda x + (1-\lambda)y \in S_i \quad i = 1, \ldots, n$

$\rightarrow \lambda x + (1-\lambda)y \in S \rightarrow S$ è convesso

Inoltre:

- Se $S = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0 \quad i = 1, 2, \ldots, n\} \quad e \quad g_i$ convessare $i = 1, \ldots, n ==> S$ è convessa
- Se $S = \{h(x) = 0\}$ è convesso $\Longleftrightarrow h(x)$ è lineare

5.5 Concavità

Funzione concava

Una funzione $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ è **concava** su S se $-f(x)$ è convessa su S. In altre parole, se e solo se:

$$f(\lambda x + (1-\lambda)y) \geqslant \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in S \quad \forall \lambda \in [0,1]$$

Una funzione $f : \mathbb{R}^n \rightarrow \mathbb{R}$ è **strettamente concava** se $-f(x)$ è strettamente convessa. In altre parole, se e solo se:

$$f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in S \quad \forall \lambda \in [0,1]$$

Una funzione lineare è sia concava che convessa.

5.6 Applicazioni convessità

Ricordiamo che il problema che vogliamo risolvere è del tipo:

$$\begin{cases} \min f(x) \\ x \in S \end{cases}$$

Ipotesi:

1. f convesso su S: $\forall x, y \in S \quad f(\lambda x) + (1-\lambda)y \leqslant \lambda f(x) + (1-\lambda) f(y) \quad \forall \lambda \in [0,1]$
   (strattamente convesso$\longrightarrow f(\lambda x) + (1-\lambda)y \leqslant \lambda f(x) + (1-\lambda) f(y)$)
2. S convesso $\forall x, y \in S \quad \lambda x + (1-\lambda)y \in S \quad \forall \lambda \in [0,1]$

Proposizione 1. *se* $\begin{cases} \min f(x) \\ x \in S \end{cases}$ *, f convessa su S e S convessa. Allora valgono le seguenti proposizioni:*

1. *Non $\exists$ minimi locali non globali; [Non si assicura l'$\exists$ della soluzione];*
2. *L'insieme delle soluzioni ottime è un insieme convesso;*

Dimostrazione. (1)

Supponiamo che $x^* \in S$ è un iminimo locale$\rightarrow \exists \rho > 0 t.c. \, f(x^*) \leqslant f(x) \, \forall x \in S \cap B(\rho, x^*)$

Considero un punto $x \in S, x \notin B(x^*, \rho)$ e quest'ultimo non deve essere un punto di minimo globale:

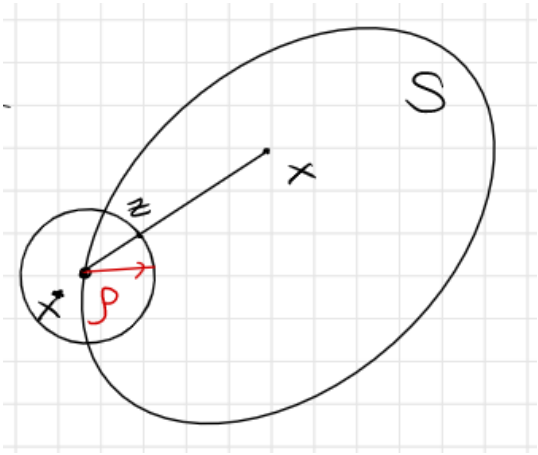


Figura 2.

In questa figura abbiamo rappresentato il nostro insieme S convesso e il punto di minimo globale x^* con il suo corrispettivo intorno. Inoltre, punto x soddisfa l'ipetesi cui sopra.

Poiché S è convesso $\rightarrow$ il segmento $[x^*, x]$ è contenuto in S. In paricolare è tale che:

$$\lambda x + (1-\lambda)x^* \in S \quad \forall \lambda \in [0,1]$$

Il punto z è quel punto che si trova nell'intorno di x^* e sul segmento che connette $[x^*, x]$. Allora:

$$z = \bar{\lambda} x + (1-\bar{\lambda})x^* \in S \cap B(\rho, x^*)$$

Inoltre poiché x^* è pun punto di minimo:

$$f(x^*) \leqslant f(z) \qquad \text{dato che } z \cap B(\rho, x^*)$$

$$f(x^*) \leqslant f(\bar{\lambda}x + (1-\bar{\lambda})x^*) \leqslant \bar{\lambda} f(x) + (1-\bar{\lambda}) f(x^*) \text{ poiché } f \text{ convessa}$$

Quindi:

$$f(x^*) \leqslant \bar{\lambda} f(x) + f(x^*) - \bar{\lambda} f(x^*)$$

$$f(x^*) \leqslant f(x) \quad \forall x \in S$$

Quindi abbiamo dimostrato che non se il problema è di minimizzazione, la funzione e l'insieme sono convessi, allora non esiste un minmo locale che non è globale.

Inoltre, se considerassinmo la convessità stretta otterremmo che x^* è l'**unico** punto di minimo globale. □

Dimostrazione. (2)

Se l'insieme S è convesso, definiamo $X^* := \{\text{insieme soluzioni globali}\}$. I casi di interesse sono i seguenti:

- $X^* = \varnothing \rightarrow$ Insieme convesso;
- $X^* = \{x^*\} \rightarrow$ Insieme convesso;
- $|X^*| \geqslant 2$ ci sono almeno due soluzoni globali; □

Nell'ultimo caso, supponiamo che $x_1^* \in X^*, x_2^* \in X^*$ se sono soluzioi globali allora:

$$x_1^* \in S \rightarrow f(x_1^*) \leqslant f(x) \quad \forall x \in S$$

$$x_2^* \in S \longrightarrow f(x_2^*) \leqslant f(x) \quad \forall x \in S$$

Necessariamente se sono **punti di minimo globali** deve valere:

$$f(x_1^*) = f(x_2^*) = f^*$$

Inoltre S è convesso$\rightarrow \forall \lambda \in [0,1] \quad \lambda x_1^* + (1-\lambda) x_2^* \in S$

Poiché f convessa: $f(\lambda x_1^* + (1-\lambda)x_2^*) \leqslant \lambda f(x_1^*) + (1-\lambda) f(x_2^*) = \lambda f^* + f^* - \lambda f^*$

$$f(\lambda x_1^* + (1-\lambda)x_2^*) \leqslant f^*$$

Poiché f^* è ottimo globale per ipotesi$\rightarrow \lambda x_1^* + (1-\lambda)x_2^* \in X^* \rightarrow X^*$ è un insieme convesso.

5.7 Derivata direzionale di f in x lungo direzion d

In questo corso consideriamo funzioni $f \in C^1$ che sono **continuamente differenziabili** e le sue derivate sono **funzioni continue**.

Definizione 2. *Una **derivata direzionale** $Df(x,d)$ se esiste:*

$$\lim_{t \rightarrow 0} \frac{f(x + t\,d) - f(x)}{t} = \mathrm{Df}(x,d)$$

Definizione 3. *Definiamo la **derivata parziale i-esima in d**:*

$$D\,f(x, e_i) \quad d = e_i = \begin{pmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{pmatrix} \leftarrow i - \text{esima componente}$$

$$\frac{\partial f(x)}{\partial x_i} = \lim_{t \rightarrow 0} \frac{f(x_1, \ldots, x_i + t, \ldots, x_n) - f(x_i)}{t}$$

Definizione 4. ***Gradiente di $f(x)$***

Il gradiente di $f(x)$ è un vettore colonna nella forma:

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

Corollario 5.

Se $f \in C^1$ su $S \rightarrow D\,f(x,d) = \nabla f(x)^T d = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} = \sum_{i=1}^n \frac{\partial f(x)}{\partial x_i} d_i$

Nota 6.

$$x^T y = \sum_{i=1}^{n} x_i y_i = \|x\| \, \|y\| \cos(\theta) \rightarrow x^T y \leqslant \|x\| \, \|y\|$$

5.7.1 Sviluppo di Taylor I ordine

$$f(x+d) = f(x) + \nabla f(x)^T d + \alpha(x,d)$$

Il resto $\alpha\,(x,d)$ è **trascurabile** poiché:

$$\lim_{\|d\| \downarrow 0} \frac{\alpha(x,d)}{\|d\|} = 0$$

Quindi se $\|d\|$ è sufficientemente piccolo $\rightarrow f(x+d) \approx f(x) + \nabla f(x)^T d$ è **lineare**

Proposizione 7. *Convessità e derivara del primo ordine*

$f \in C^1$ *su* S *e* f *è convessa su* $S \iff \forall x, y \in S \quad f(y) \geqslant f(x) + \nabla f(x)^T (y-x)$

Inoltre:

$f \in C^1 \,\text{su}\, S$ *e* f *strettamente convessa su* $S \iff \forall x, y \in S \quad f(y) > f(x) + \nabla f(x)^T (y-x)$

5.7.2 Sviluppo Taylor II ordine

Definizione 8. ***Matrice Hessiana***

Una matrice $\nabla^2 f(x)$ *è una* ***matrice hessiana*** *se è simmetrica ed è nella seguente forma:*

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

Per esistere, $f \in C^2$

A questo punto possiamo introdurre lo **sviluppo al II ordine di Taylor:**

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \beta(x,d)$$

Questo sviluppo è migliore dello sviluppo del primo ordine poiché tende a zero più velocemente. Infatti:

$$\lim_{\|d\| \downarrow 0} \frac{\beta(x,d)}{\|d\|^2} = 0$$

In pratica è più accurato in uno sppazio più grande.

5.8 Segno di una matrice

Definizione 9. *(Semi)definita positiva/negativa*

Data una matrice $Q \in \mathbb{R}^{n \times n}$, $Q = Q^T$. *Definiamo:*

- Q ***semidefinita positiva:*** $Q \succcurlyeq 0 \iff y^T Q\, y \geqslant 0 \quad \forall y \in \mathbb{R}^n$

- Q ***definita positiva:*** $Q \succ 0 \iff y^T Q\, y > 0 \quad \forall y \in \mathbb{R}^n$
- Q ***definita negativa:*** $Q \prec 0 \iff y^T Q\, y < 0 \quad \forall y \in \mathbb{R}^n$
- Q ***semidefinita negativa:*** $Q \preccurlyeq 0 \iff y^T Q\, y \leqslant 0 \quad \forall y \in \mathbb{R}^n$
- Q ***indefinita*** $\iff \exists y_1 \in \mathbb{R}^n : y_1^T Q\, y_1 > 0 \wedge \exists y_2 \in \mathbb{R}^n : y_2^T Q\, y_2 < 0$

5.8.1 Matrice Hessiana

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

Proposizione 10. *$f(x)$ convessa su* $\forall x, y \in S \quad \lambda \in [0,1] \quad f(\lambda x + (1-\lambda) y) \leqslant \lambda f(x) + (1-\lambda) f(y)$

Proposizione 11. *$+ f \in C^2$ e f convessa su $S \iff \nabla^2 f(x) \succcurlyeq 0 \forall x \in S$*

Nota 12.

- Se si trova 1 punti per cui non vale 11, allora non si ha la convessità

5.8.2 Criterio dei minori di Nord-Ovest

Ipotizziamo $Q \in \mathbb{R}^{n \times n}, Q = Q^T$

Definizione 13. *Minori di Nord-Ovest di ordine k*

I minori di Nord-Ovest di ordine k sono quei minori ottenuti cancellando le ultime $n-k$ righe e colonne.

Quindi se :

$$Q \succcurlyeq 0 \iff \text{tutti } i \text{ minori di nord} - \text{ovest hanno determinante} > 0 \text{ di ordine } k = 1, \ldots, n$$

Definizione 14. *Matrice diagonale dominante*

La matrice diagonale dominante è una condizione ***sufficiente*** *di definita positività. In particolare date le ipotesi 5.8.2, si ha una matrice diagonale dominante se:*

$$q_{ii} > \sum_{j \neq i} |q_{ij}| \quad \forall i$$

Nel caso in cui non è possibile determinare la semidefinita/definita positività occore considerare i **minori principali.**

Definizione 15. *Minori principali di ordine k*

*I **minori principali di ordine k** sono tutti i minori che si appoggiano sulla diagonale eliminando n-k righe e colonne in tutti i modi possibili.*

Importante 16. Algoritmo segno di una matrice

1. Applico criterio di nord-ovest:

   A) tutti i minori hanno determinante >0 $\longrightarrow$STOP $\quad Q \succ 0$

   B) Si ha un minore con determinante $=0$ $\longrightarrow$**Criterio minori principali**:

      i. tutti i minori principali hanno determinante $\geqslant 0 \longrightarrow$ STOP $\longrightarrow Q \succcurlyeq 0$

      ii. trovo minore con determinante <0:

         a) se i minori principali hanno determinante >0 e un determinante <0

            $\longrightarrow$STOP $\Longrightarrow Q$ indefinita

         b) Ritorno al passo 1 considerando $-Q$ per dedurne la (semi)definita negatività.

5.8.3 Segno di una matrice - Autovalori/Autovettori

Sia data una matrice $Q = Q^T, Q \in \mathbb{R}^{n \times n}$:

Definizione 17. *Autovalore, Autovettore*

*λ_i è un **autovalore** di Q con corrispondente **autovettore** u_i se:*

$$Q\,u_i = \lambda_i\,u_i \quad i = 1, \ldots, n \quad \|u_i\| = 1$$

***Nota* 18.** *$u_i^T u_j = 0$ gli autovettori sono ortogonali per autovalori diversi $\quad \lambda_i \neq \lambda_j$*

Quindi:

$$Q = \sum_{i=1}^{n} \lambda_i u_i\,u_i^T$$

Si ha che:

- Q **semidefinita positiva** $\Longleftrightarrow \lambda_1, \ldots, \lambda_n \geqslant 0$ $(y^T Q\,y \geqslant 0 \forall y \in \mathbb{R}^n \Longleftrightarrow \lambda_{\min}(Q) \geqslant 0)$
- Q **definita positiva** $\Longleftrightarrow \lambda_1, \ldots, \lambda_n > 0$ $(y^T Q\,y > 0 \forall y \in \mathbb{R}^n \Longleftrightarrow \lambda_{\min}(Q) > 0)$
- Q **semidefinita negativa** $\Longleftrightarrow \lambda_1, \ldots, \lambda_n \leqslant 0$ $(y^T Q\,y \leqslant 0 \forall y \in \mathbb{R}^n \Longleftrightarrow \lambda_{\max}(Q) \leqslant 0)$
- Q **definita negativo** $\Longleftrightarrow \lambda_1, \ldots, \lambda_n < 0$ $(y^T Q\,y < 0 \forall y \in \mathbb{R}^n \Longleftrightarrow \lambda_{\max}(Q) < 0)$
- Q **indefinita** $\Longleftrightarrow \exists \lambda_i > 0, \lambda_j < 0$

$$\lambda_{\min}(Q) = \min_{x \in \mathbb{R}^n} \frac{x^T Q\,x}{\|x\|^2} := \textbf{quoziente di Rayleigh}$$

La seguente formula permette di valurare la coercività:

$$x^T Q\, x \geqslant \lambda_{\min}(Q)\|x\|^2$$

6 Condizioni di esistenza delle soluzioni

Consideriamo il problema nella forma:

$$\begin{cases} \min f(x) \\ x \in S \end{cases}$$

6.1 Definizioni preliminari

Definizione 19. *Insieme aperto*

*Un insieme S è **aperto** se $\forall x \in S, \exists B(x, \rho)$ aperta tutta contenuta in S.*

Definizione 20. *Insieme chiuso*

*Un insieme S è **chiuse** se il suo complemento $\mathbb{R}^n / S$ è aperto.*

Definizione 21. *Insieme limitato*

*Un insieme S è **limitato** se $\exists M > 0 : \forall x \in S \qquad \|x\| \leqslant M$.*

Definizione 22. *Insieme compatto*

*Un insieme S è **compatto** se è **chiuso** e **limitato**.*

6.2 Esistenza del minimo

Teorema 23. ***Teorema di Weierstrass***

Se vogliamo $\begin{cases} \min f(x) \\ x \in S \end{cases}$ e f continua e S compatta $\longrightarrow \exists$ minimo (massimo) globale.

Importante 24. Il teorema di Weierstrass fornisce condizioni solo sufficienti: se non è verificata non si può concludere nulla.

L'ipotesi di compattezza di S può essere rilassata, imponendo una condizione più stringente sulla f.

Definizione 25. *Insieme di Livello vincolato di f*

Dato il problema di minimizzazione:

$$\begin{cases} \min f(x) \\ x \in S \end{cases}$$

*Si definisce **insieme di livello vincolato di f**:*

$$\mathcal{L}(\alpha) = \{x \in S : f(x) \leqslant \alpha\} \quad \alpha = f(x_0), x_0 \in S \text{ iniziale}$$

Teorema 26.

Se $\exists x_0 \colon \mathcal{L}(f(x_0))$ *compatto e* f *continua* $\Longrightarrow$ $\exists$una soluzione globale

Dimostrazione.

Se si considera il problema di minimizzazione $\begin{cases} \min f(x) \\ x \in \mathcal{L}(\alpha) \end{cases}$ *, vale il teorema di Weierstrass.*

Quindi esiste una soluzione globale x^* *tale che* $f(x^*) \leqslant f(x) \quad \forall x \in \mathcal{L}(f(x_0))$.

Se:

$x \notin \mathcal{L}(f(x_0)), \text{ma} \quad x \in S \rightarrow f(x) > f(x_0) \geqslant f(x^*) \rightarrow f(x^*) \leqslant f(x) \forall x \in \mathcal{L}(f(x_0)), \forall x \in S \colon x \notin \mathcal{L}(f(x_0)) \Longrightarrow$

$f(x^*) \leqslant f(x) \quad \forall x \in S \longrightarrow x^*$ è un minimo globale. □

Definizione 27. *Coercività di f*

La funzione f si dice ***coerciva*** *su S se comunque scelgo* $\{x_k\} \in S$ *tale che:*

$\lim_{k\to\infty} \|x_k\| = \infty$ *e* $\lim_{k\to\infty} f(x_k) = \infty$

Teorema 28.

f è ***coerciva*** *su S* $\iff$ *tutti i suoi insiemi di livello sono compatti*

Teorema 29. *Esistenza del minimo (Sufficiente)*

Se si considera il problema di minimizzazione $\begin{cases} \min f(x) \\ x \in S \end{cases}$ *e la funzione f è continua e coerciva su S. Allora esiste la soluzione di minimo globale.*

Importante 30.

Se $Q \succ 0$ ed $f(x)$ è strettamente convessa allora la soluzione esiste ed è unica su qualunque S chiuso e non vuoto.

Definizione 31. *Termine di regolarizzazione*

Sia dato il problema di minimizzazione $\begin{cases} \min f(x) \\ x \in S \end{cases} \longrightarrow \min f(x) + \tau \|x - x_k\|^2 \quad \tau >' 0$

Nota 32.

- Il termine $\tau \|x - x_k\|$ rende coerciva la funzione e la convessifica rendendo il problema più semplice. Poiché la matrice Hessiana diventa costante.

7 Condizioni di ottimo per problemi non vincolati

Con riferimento ad un problema di minimizzazione del tipo:

$$\min f(x), \quad x \in S$$

una **condizione di ottimalità** è una condizione (necessaria,sufficiente,necessaria e sufficiente) perché un punto x^* risulti una soluzione ottima locale o globale del problema.

Importante 33. Condizioni necessarie

Una **condizione necessaria** può servire a restringre l'insieme in cui ricercare le soluzioni del problema necessario e a costruire algoritmi che soddisfino tali condizioni:

$$x^* \text{ minimo locale} \in S \longrightarrow \text{vale la condizione}$$

Quindi forniscono **criteri di arresto** per gli algoritmi di risoluzioni e una indicazione su dove cercare l'ottimo.

Importante 34. Condizioni sufficiente

Una **condizione sufficiente** può servere a dimostrare che un punto ottenuto sia una soluzione ottima, locale o globale, del problema.

$$\text{vale la condizione} \rightarrow x^* \text{ minimo locale} \in S$$

Importante 35. Condizioni necessarie e sufficienti

Questo tipo di condizioni si hanno solamente nelle tipologie di problemi convessi.

$$x^* \text{ minimo globale} \longleftrightarrow \text{vale condizione}$$

Da queste considerazioni possiamo trarne una rappresentazione insiemistica:

Da cui possiamo evincere che, se la condizione necessaria è violata, il punto può essere escluso e che le condizioni suddicienti sono utilisi sole soddisfatte; in alternativa non si può concludere nulla dal problema.

7.0.1 Direzione di discesa

L'idea è quella di definire una **direzione di discesa** in un punto x come un vettore $d \in \mathbb{R}^n$ tale che, per tutti gli spostamenti sufficientemente piccoli lungo d, si ha una diminuzione stretta del valore di f rispetto al valore di x

Definizione 36. *Direzione di discesa*

*Sia $f: R^n \rightarrow R$ e $x \in \mathbb{R}^n$. Si dice che un vettore $d \in \mathbb{R}^n, d \neq 0$ è una **direzione di discesa** per f in x se esiste $\bar{t}>0$ tale che:*

$$f(x + t\,d) < f(x) \qquad \forall t \in (0, \bar{t}]$$

In modo analgo si potrebbe definire una direzione di salita nel punto x se f aumenta strettamente per tutti gli spostamenti abbastanza piccoli lungo d

La direzione di discesa la si può caratterizzare tramite le derivate direzionali della funzione obiettivo. Infatti ricordiamo che la **derivata direzionale** è definita nel seguente modo:

$$D\,f(x, d) = \lim_{t \rightarrow 0} \frac{f(x + d\,t) - f(x)}{t}$$

Proposizione 37. *Derivata direzionale e direzione di discesa*

Se $f \in C^1$ continua e derivabile almeno una volta$\longrightarrow D f(x,d) = \nabla f(x)^T d$. Quindi, per t sufficientemente piccolo,otteniamo:

1. $\nabla f(x)^T d < 0 \longrightarrow f(x+t\,d) - f(x) < 0 \longrightarrow d$ *è di* ***discesa****:*
2. $\nabla f(x)^T d > 0 \longrightarrow f(x+t\,d) - f(x) > 0 \longrightarrow d$ *è di* ***salita****;*
3. $\nabla f(x)^T d = 0 \longrightarrow f(x+t\,d) - f(x) = 0 \longrightarrow$ *?????*

Teorema 38. ***Condizione di discesa del primo ordine*** *(necessaria e sufficiente)*

Se f convessa e C^1 ,d discesa $\Longleftrightarrow \nabla f(x)^T d < 0$

Dimostrazione.

Prima implicazione:($\Longleftarrow$)

Se $\nabla f(x)^T d < 0$, la direzione d è di discesa in quanto, per la definizione di derivata direzionale, per valori positivi sufficientemente piccoli di t, deve valere $f(x+d\,t) - f(x) < 0$.

Seconda Implicazione: ($\Longrightarrow$)

Poniamo $y = x + t\,d$ ed assumiamo che f sia convessa. Dalla definizione di convessità:

$$\begin{aligned} f(y) &\geqslant f(x) + \nabla f(x)^T (y-x) \\ f(x+t\,d) &\geqslant f(x) + t\nabla f(x)^T (y-x) \\ &\geqslant f(x) + t\,\nabla f(x)^T d \geqslant f(x) \end{aligned}$$

per cui $\nabla f(x)^T d \geqslant 0$ implica $f(x+t\,d) \geqslant f(x)$. Tuttavia, se d è una direzione di discesa, deve necessariamente essere $\nabla f(x)^T d < 0$. □

Definizione 39. *Segno derivata direzionale e antigradiente*

Dal punto di vista geometrico, ricordando il concetto di angolo fra due vettori possiamo studiare il ***segno derivata direzionale****:*

$$\nabla f(x)^T d = \|\nabla f(x)\|\,\|d\| \cos(\theta)$$

Se:

— $\theta > 90° \longrightarrow \cos(\theta) < 0 \longrightarrow \nabla f(x)^T d < 0 \longrightarrow d$ *è di* ***discesa****;*

— $\theta < 90° \longrightarrow \cos(\theta) > 0 \longrightarrow \nabla f(x)^T d > 0 \longrightarrow d$ *è di* ***salità***

— *Il caso "limite", cioè di* ***massima discesa*** *si ottiene con il cosiddetto* ***antigradiente****:*

$$d = -\nabla f(x) \longrightarrow \theta = 180° \longrightarrow \cos(\theta) = 1 \text{ è sempre di } \textbf{discesa}$$

$$\nabla f(x)^T d = -\|\nabla f(x)\|^2 < 0$$

7.0.2 Direzioni a curvatura negativa in x

Se f è differenziabile due volte è possibile caratterizzare l'andamento di f lungo una direzione assegnata utilizzando anche le derivate seconde per stabilire le condizioni di ottimo del secondo ordine.

Definizione 40. *Direzione a curvatura negativa*

Sia $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ *due volte continuamente differenziabile nell'intorno di un punto* $x \in \mathbb{R}^n$. *Si dice che un vettore* $d \in \mathbb{R}^n, d \neq 0$ *è una* ***direzione a curvatura negativa*** *per* f *in* x *se risulta:*

$$d^T \nabla^2 f(x)\, d < 0$$

Nota 41. Una direzione a curvatura negativa è quindi tale che la derivata direzionale seconda è negativa in x, per cui diminuisce localmente la derivata direzionale del primo ordine.

Teorema 42. ***Condizione di discesa del secondo ordine***

Sia $f: \mathbb{R}^n \to \mathbb{R}$ *due volte continuamente differenziabile nell'intorno di un punto* $x \in \mathbb{R}^n$ *e dia* $d \in \mathbb{R}^n \neq 0$. *Supponiamo che risulti* $\nabla f(x)^T d = 0$, *e che* d *sia una direzione a curvatura negativa in* x, *ossia tale che* $d^T \nabla^2 f(x)\, d < 0$. *Allora* d *è una* ***direzione di discesa*** *per* f *in* x.

Dimostrazione.

Poiché f è differenziabile due volte, si ha:

$$f(x + t\, d) = f(x) + t \nabla f(x)\, d + \frac{1}{2} t^2 d^T \nabla^2 f(x)\, d + \beta(x, t\, d)$$

In cui β è il resto:

$$\lim_{t \to 0} \frac{\beta(x, t\, d)}{t^2} = 0$$

Essendo per ipotesi $\nabla f(x)\, d = 0$, si può scrivere:

$$\frac{f(x + t\, d) - f(x)}{t^2} = \frac{1}{2} d^T \nabla^2 f(x)\, d + \frac{\beta(x, t\, d)}{t^2}$$

Poichè $d^T \nabla^2 f(x)\, d < 0$ per definizione di curvatura negativa e $\frac{\beta(x,\, t\, d)}{t^2} \to 0 \quad \text{per } t \to 0$. Per valori sufficientemente piccoli di t, si ha che $f(x + t\, d) - f(x) < 0$ e quindi d è una direzione di discesa. □

Definizione 43. *Segno derivata direzionale e antigradiente*

Dal punto di vista geometrico, ricordando il concetto di angolo fra due vettori possiamo studiare il ***segno derivata direzionale****:*

$$\nabla f(x)^T d = \|\nabla f(x)\|\, \|d\| \cos(\theta)$$

Se:

— $\theta > 90° \longrightarrow \cos(\theta) < 0 \longrightarrow \nabla f(x)^T d < 0 \longrightarrow d$ *è di* ***discesa****;*

— $\theta < 90° \longrightarrow \cos(\theta) > 0 \longrightarrow \nabla f(x)^T d > 0 \longrightarrow d$ *è di* ***salità***

— *Il caso "limite", cioè di* ***massima discesa*** *si ottiene con il cosiddetto* ***antigradiente****:*

$$d = -\nabla f(x) \longrightarrow \theta = 180° \longrightarrow \cos(\theta) = 1 \text{ è sempre di } \textbf{discesa}$$

$$\nabla f(x)^T d = -\|\nabla f(x)\|^2 < 0$$

7.0.3 Direzioni a curvatura negativa in x

Se f è differenziabile due volte è possibile caratterizzare l'andamento di f lungo una direzione assegnata utilizzando anche le derivate seconde per stabilire le condizioni di ottimo del secondo ordine.

Definizione 44. *Direzione a curvatura negativa*

Sia $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ due volte continuamente differenziabile nell'intorno di un punto $x \in \mathbb{R}^n$. Si dice che un vettore $d \in \mathbb{R}^n, d \neq 0$ è una ***direzione a curvatura negativa*** *per f in x se risulta:*

$$d^T \nabla^2 f(x)\, d < 0$$

Nota 45. Una direzione a curvatura negativa è quindi tale che la derivata direzionale seconda è negativa in x, per cui diminuisce localmente la derivata direzionale del primo ordine.

Teorema 46. ***Condizione di discesa del secondo ordine***

Sia $f: \mathbb{R}^n \to \mathbb{R}$ due volte continuamente differenziabile nell'intorno di un punto $x \in \mathbb{R}^n$ e dia $d \in \mathbb{R}^n \neq 0$. Supponiamo che risulti $\nabla f(x)^T d = 0$, e che d sia una direzione a curvatura negativa in x, ossia tale che $d^T \nabla^2 f(x)\, d < 0$. Allora d è una ***direzione di discesa*** *per f in x.*

Dimostrazione.

Poiché f è differenziabile due volte, si ha:

$$f(x + t\,d) = f(x) + t \nabla f(x)\, d + \frac{1}{2} t^2 d^T \nabla^2 f(x)\, d + \beta(x, t\,d)$$

In cui β è il resto:

$$\lim_{t \to 0} \frac{\beta(x, t\,d)}{t^2} = 0$$

Essendo per ipotesi $\nabla f(x)\, d = 0$, si può scrivere:

$$\frac{f(x + t\,d) - f(x)}{t^2} = \frac{1}{2} d^T \nabla^2 f(x)\, d + \frac{\beta(x, t\,d)}{t^2}$$

Poichè $d^T \nabla^2 f(x)\, d < 0$ per definizione di curvatura negativa e $\frac{\beta(x, t\,d)}{t^2} \to 0 \quad \text{per } t \to 0$. Per valori sufficientemente piccoli di t, si ha che $f(x + t\,d) - f(x) < 0$ e quindi d è una direzione di discesa. □

8 Condizioni di ottimo per problemi vincolati con insieme ammissibile

Il problema ce vogliamo risolvere è del tipo:

$$\begin{aligned} \min \quad & f(x) \\ & x \in S \end{aligned}$$

cercando le direzioni di discesa ammissibile del problema, cioé che non violano le condizioni dettate da $x \in S$.

Definizione 47. *Direzione ammissibile*

Sia S un sottoinsieme di $\mathbb{R}^n$ e $x \in S$. Si dice che un vettore $d \in \mathbb{R}^n, d \neq 0$ è una ***direzione ammissibile*** *per S in s se:*

$$d \in \mathbb{R}^n : \exists \bar{t} > 0 \quad x + t\,d \in S \quad \forall t \in (0, \bar{t}]$$

Proposizione 48. *Condizione di ottimo necessaria di minimo locale*

Sia $x^* \in S$ un punto di minimo locale del problema:

$$\min f(x), x \in S;$$

allora non può esistere una direzione ammissibile in x^* che sia anche di discesa.

Dimostrazione.

Per assurdo, se esistesse una direzione d che sia ammissibile e di discesa in x^*, allora in ogni intorno di x^* sarebbe possibile trovare un punto $x^* + t\,d \in B(x^*, \rho)$ tale che:

$$f(x^* + t\,d) \leqslant f(x^*)$$

il che contraddisce l'ipotesi che x^* sia un punto di minimo globale. □

Se f è differenziabile la condizione:

$$\nabla f(x)^T d < 0$$

è una condizione sufficiente perché d sia una direzione di discesa per f in x, otteniamo la condizione encessaria seguente.

Proposizione 49. *Condizione necessaria del primo ordine*

Supponiamo che $f : \mathbb{R}^n \to \mathbb{R}$ sia continuamente differenziabile nell'intorno in un punto di minimo locale $x^* \in S$ del problema:

$$\min f(x) \quad x \in S$$

Allora non può essitere una direzione ammissibile in d in x^* tale che:

$$\nabla f(x^*)^T d < 0$$

e quindi si ha:

$$\nabla f(x^*)^T d \geqslant 0 \quad \forall d \in \mathbb{R}^n \text{ ammissibile din } x^*$$

Proposizione 50. *Condizione necessaria del secondo ordine*

Supponiamo che $f : \mathbb{R}^n \to \mathbb{R}$ sia due volte continuamente differenziiabile nell'intorno di un punto di minimo locale $x^\star \in S$ del problema:

$$\min f(x) \quad x \in S$$

Allora non può esistere una direzione ammissibile d in x^* tale che:

$$\nabla f(x^*)^T d = 0 \quad d^T \nabla^2 f(x)\, d < 0$$

e quindi si ha:

$$\forall d \text{ ammissibile tale che } \nabla f(x^*)^T d = 0 \quad d^T \nabla^2 f(x)\, d > 0$$

***Nota* 51.**

Non si ha l'uguaglianza poiché avremmo una direzione a curvatura negativa e quindi ammissibile nell'insieme.

8.1 Condizioni di ottimo per problemi vincolati con insieme ammissibile convesso

Se l'insieme ammissibile S è convesso, vogliamo caratterizzare le condizioni di ottimo. A tale scopo occorre caratterizzare le direzioni ammissibili.

Proposizione 52. *Direzione ammissibile (S convesso)*

$$\begin{matrix} S \text{ convesso} \\ \\ d \text{ ammissibile in } x \end{matrix} \quad \Longleftrightarrow \quad d = y - x \quad \text{per qualche } y \in S$$

Dimostrazione.

Sia $x \in S$ allora comunque si fissi $y \in S$ tale che $x \neq y$, per la convessità di S si ha che:

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

$$x + \lambda(y-x) \in S \longrightarrow d = y - x \in S \text{ ammissibile}$$

Al contrario se $d \in \mathbb{R}^n \neq 0$ è una direzione ammissibile per S,in x esiste un punto $y \in S$ ed uno scalare $\lambda > 0$ tali che $d = \lambda(y-x)$. □

Proposizione 53. *Condizione necessaria di minimo locale del primo ordine(S convesso)*

Sia $x^* \in S$ un punto di minimo locale del problema:

$$\min f(x) \quad x \in S$$

in cui $S \subseteq \mathbb{R}^n$ è un insieme convesso e supponiamo f sia continuamente differenziabile in un intorno di x^*. Allora si ha necessariamente:

$$\nabla f(x^*)^T (y - x^*) \geqslant 0 \quad \forall y \in S$$

Proposizione 54. *Condizione necessaria di minimo locale del secondo ordine (S convesso)*

Sia $x^* \in S$ un punto di minimo locale del problema:

$$\min f(x) \quad x \in S$$

in cui $S \subseteq \mathbb{R}^n$ è un insieme convesso e supponiamo f sia due volte continuamente differenziabile in un intorno di x^*. Allora si ha necessariamente:

$$\forall y \in S \quad \nabla f(x^*)\,(y - x^*) = 0 \rightarrow (y - x^*)\, \nabla^2 f(x)\, (y - x^*) \geqslant 0$$

Questo si riconduce alla ricerca dei **punti di stazionarietà** della funzione. Infatti:

- Se il punto è interno all'insieme S:

$$\nabla f(x^*)^T (y - x^*) \geqslant 0 \quad \forall y \in S$$

$$\Updownarrow$$

$$\nabla f(x^*) = 0$$

  Questa condizione è detta di **stazionarietà non vincolata.**

- Se ci troviamo sul bordo dell'insieme S:

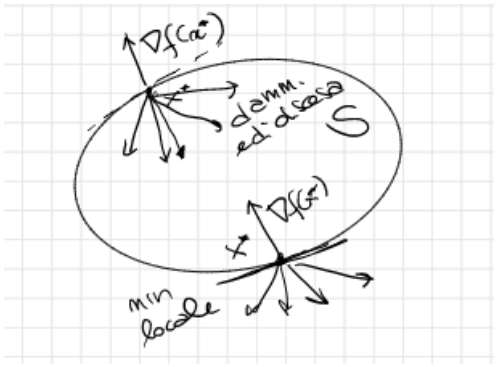


Figura 3.

Dipende poiché tutte le direzioni di discesa sono non ammissibili. In alto, se il cos>90 otteniamo direazioni ammissibili di discesa e non abbiamo minimo; in basso, se il cos>90 le direzioni di discesa non sono ammissibile ed otteniamo un minimo locale.

Dimostrazione.

$\Longrightarrow$) x^* minimo globale $\rightarrow x^*$ minimo locale $\rightarrow$ vale la C.N. del primo ordine;

$\Longleftarrow$) Poiché f convessa su $S \quad \forall y \in S$

$$f(y) \geqslant f(x^*) + \nabla f(x)^T (y - x^*) \geqslant f(x^*)$$

In cui $\nabla f(x)^T (y - x^*) \geqslant 0$ per ipotesi. Allora:

$$f(x^*) \leqslant f(y) \quad \forall y \in S \longrightarrow x^* \text{ minimo globale} \qquad \square$$

Importante 55.

Se f strettamente convessa$\longrightarrow$minimo globale diventa unico

8.2 Problemi con vincoli lineari

$$\begin{array}{ll} \min & f(x) \qquad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \\ & Ax \leqslant b \end{array}$$

Inoltre:

$$a_i^T x \leqslant b_i \quad i = 1, \ldots, m; a^T \text{ righe di } A$$

Supponiamo $\overline{x} \in S \longleftrightarrow a_i^T \overline{x} \leqslant b_i \quad i = 1, \ldots, m$

Definizione 56. *Vincoli attivi nel punto*

$$I(x) := \{i = 1, \ldots, m : a_i^T x = b\}$$

***Nota* 57.**

I vincoli attivi influenzano le direzioni ammissibili

Proposizione 58. *Direzioni ammissibili: disuguaglianze lineari*

*Sia x^* tale che $a_i^T x^* \leqslant b_i$ con $i = 1, \ldots, m$. Allora un vettore non nullo $d \in \mathbb{R}^n$ è una **direzione ammissibile** per S se e solo se:*

$$a_i^T d \leqslant 0 \quad \forall i \in I(x^*)$$

Dimostrazione.

$\Longrightarrow$)Dati un vettore non nullo $d \in \mathbb{R}^n$ e numero $t \geqslant 0$ qualsiasi sufficientemente piccolo, il punto $x^* + t\,d$ è ammissibile relativamente ai vincoli attivi in x^* se e solo se:

$$a_i^T (x^* + t\,d) = b_i + t\,a_i^T d \leqslant b_i \qquad \forall i \in I(x^*)$$

Quindi se e solo se vale la proposizione 58.

$\Longleftarrow$)Se vale $a_i^T d \leqslant 0 \quad \forall i \in I(x^*)$, d è una direzione ammissibile anche per i vincoli non attivi. Infatti, essendo $a_i^T x^* \leqslant b_i$ per $i \notin I(x^*)$ si può assumere t sufficientemente piccolo da avere:

$$a_i^T (x^* + t\,d) \leqslant b_i \quad \forall i \notin I(x^*)$$

Quindi si può concludere che esiste un numero $\bar{t} > 0$ tale che $A(x^* + t\,d) \leqslant b \quad \forall t \in [0, \bar{t}]$, ossia che d è una direzione ammissibile per S in x^*, □

Importante 59. L'insieme delle direzioni ammissibili è un **poliedro**.

Importante 60. Se ho vincoli soffisfatti all'uguaglianza la disequazione 58 viene soddisfatta all'uguaglianza.

8.3 Condizioni di ottimalità per problemi con vincoli di box

Sia $x^* \in \mathbb{R}^n$ un punto di minimo locale del problema

$$\begin{cases} \min f(x) \\ l \leqslant x \leqslant u \end{cases}$$

In cui $l_i < u_i \quad i = 1, \ldots, n$. Allora, la condizione di stazionarietà del primo ordine è utile per calcolare il minimo locale x^*; quindi, risulta:

$$\frac{\partial f}{\partial x_i} = \begin{cases} \geqslant 0 & \text{se } x_i^* = l_i \\ = 0 & \text{se } l_i < x_i^* < u_i \\ \leqslant 0 & \text{se } x_i^* = u_i \end{cases}$$

Nota 61. L'insieme appena descritto è un insieme chiuso e limitato, quindi **compatto**.

Grazie al **teorema di Weierstrass** 23 possiamo dedurre l'esistenza della soluzione.

A questo punto, dobbiamo discutere tre casistiche:

A) Se $x_i^* = l_i$, consideriamo la direzione $d = e_i$, in cui e_i è l'i-esimo asse coordinato.Ricordiamo che:

Il vincolo è attivo $x_i^* = l_i \iff a_i^T x \leqslant b_i \longrightarrow d$ ammissibile $\iff a_i^T d \leqslant 0$

Quindi segue necessariamente che l'unica direzione ammissibile è data da:

$$\nabla f(x^*)^T d = \frac{\partial f(x^*)}{\partial x_i} \geqslant 0$$

Figura 4. direzioni ammissibili

B) Se $l_i < x_i^* < u_i$ le direzioni $d = e_i, d = -e_i$ sono ammissibili e dalla proposizione si ottiene:

$$\nabla f(x^*)^T d = \frac{\partial f(x^*)}{\partial x_i} \geqslant 0; \nabla f(x^*)^T d = -\frac{\partial f(x^*)}{\partial x_i} \geqslant 0$$

Da cui:

$$\nabla f(x^*)^T d = \frac{\partial f(x^*)}{\partial x_i} = 0$$

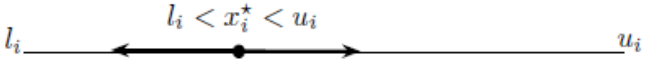


Figura 5. Direzioni ammissibili

C) Se $x_i^* = u_i$, si ha la direzione $d = -e_i$ ammissibile per cui otteniamo che:

$$\nabla f(x^*)^T d = \frac{\partial f(x^*)}{\partial x_i} \leqslant 0$$

Figura 6.

9 Condizioni di Ottimo per problemi vincolati

Consideriamo i problemi con vincoli di eguaglianza e diseguaglianza del tipo:

$$\begin{cases} \min f(x) \\ g_i(x) \leqslant 0 \quad i=1,\dots,n \\ h_j(x) = 0 \quad j=1,\dots,p \end{cases}$$

9.1 Condizioni di ottimo analitiche

9.1.1 Lagrangiana generalizzata

Prima di tutto vogliamo espiremere le condizioni di ottimo come estensione della *regole dei moltiplicatori di Lagrange*. A tale scopo, introduciamo i seguenti moltiplicatori:

1. λ_0 associato alla funzione obiettivo;
2. λ_i associato ai vincoli di diseguaglianza;
3. μ_j associato ai vincoli di eguaglianza;

Così facendo, possiamo definire la **Lagrangiana generalizzata**:

$$L(x,\lambda_0,\lambda,\mu) = \lambda_0 f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$$

9.1.2 Condizioni necessarie di Fritz John

Teorema 62. *Condizioni necessarie di Fritz John*

Sia x^* un punto di minimo locale del problema:

$$\begin{cases} \min f(x) \\ g(x) \leqslant 0 \\ h(x) = 0 \end{cases}$$

e supponiamo che f, g, h siano continuamente differenziabili in un interno di x^*. Allora esistono moltiplicatori $\lambda_0^* \in \mathbb{R}, \lambda^* \in \mathbb{R}^m, \mu^* \in \mathbb{R}^p$ tali che:

$$\begin{aligned} \nabla_x L(x^*,\lambda_0^*,\lambda^*,\mu^*) = 0 &= \lambda_0^* f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{j=1}^{p} \mu_j^* h_j(x^*) := \textbf{Annullamento del gradiente} \\ \lambda_i^* g_i(x^*) &= 0 \quad i=1,\dots,m \quad := \textbf{Condizione di complementarietà} \\ \lambda_0^*, \lambda^* &\geqslant 0; (\lambda_0^*, \lambda^*, \mu^*) \neq 0 := \textbf{Condizione di non negatività} \\ g_i(x^*) &\leqslant 0, h_j(x^*) = 0 := \textbf{Condizione di ammissibilità} \end{aligned}$$

9.2 Condizioni di Qualificazione dei Vincoli o Condiziondi di Regolarità

Dalla **condizone di negatività** si hanno due possibilità:

$$\lambda_0^* \geqslant 0 \rightarrow \begin{cases} \lambda_0^* = 0 \\ \lambda_0^* > 0 \end{cases}$$

Nel caso in cui $\lambda_0^* = 0 \longrightarrow \exists \lambda^*, \mu^*$ t.c.:

$$\sum_{i=1}^{m} \lambda_i^* \, g_i(x^*) + \sum_{j=1}^{p} \mu_j^* \, h_j(x^*) = 0$$

$$\lambda^* \geqslant 0, (\lambda^*, \mu^*) \neq 0$$

$$g_i(x^*) \leqslant 0 \quad h_j(x^*) = 0$$

Quindi la funzione oviettivo non è considerata: un qualunque punto x^* che soddisfa le condizioni di Fritz John è un candidato all'ottimo indipendentemente da $f \longrightarrow$ è un punto particolare dell'insieme ammissibile: un **punto non regolare.**

9.2.1 Condizioni di qualificazione dei vincoli

In questo ambito siamo interessati solo ai candidati per cui $\lambda_0^* \neq 0$ poiché coinvolgono la nostra funzione obiettivo.Le **condizioni di qualificazione dei vincoli** sono delle condizioni sui vincoli che definiscono l'insieme ammissibile in x^* che se se sono soddisfatte, allora $\lambda_0^* > 0$:

a) **LICQ (Linear Indipendence Constraint Qualification)**

x^*, candidato di minimo locale, soddisfa **LICQ** se per i vettori $\nabla g_i(x^*)$ con $i \in I(x^*) = \{i = 1, \ldots, n : g_i(x^*) = 0\}$ (insieme dei vincoli attivi) e $\nabla h_j(x^*) \quad j = 1, \ldots, p$ sono **linearmente indipendenti** allora $\lambda_0^* \neq 0$;

Dimostrazione. Si suppone per assurdo che $\lambda_0^* = 0 \rightarrow$ la **condizione di annullamento** della Lagrangiana diventa:

$$\sum_{i=1}^{m} \lambda_i^* \, g_i(x^*) + \sum_{j=1}^{p} \mu_j^* \, h_j(x^*) = 0$$

Dalla **complementarietà:**

$$\lambda_i^* \, g_i(x^*) = 0 \quad i = 1, \ldots, m$$

Quindi, se $i \notin I(x^*) \Longleftrightarrow g_i(x^*) < 0 \longrightarrow \lambda^* = 0, \mu^* = 0$

Non viene soddisfatta la complementarietà.

Inoltre, se vale LICQ$\longrightarrow \nabla g_i(x^*) \quad i \in I(x^*)$ e $\nabla h_j(x^*)$ sono <u>linearmente indipendenti</u>. Quindi l'unica soluzione è:

$$d_i = 0, i \in I(x^*) \, \mu^* = 0 \quad j = 1, \ldots, p \quad \textbf{ASSURDO}$$

b) **Linearità dei vincoli di eguaglianza e concavità dei vincoli attivi in x^***

Supponiamo che i vincoli di eguaglianza siano *lineari* e che i vincoli di diseguaglianza attivi siano concavi nel punto x^*. In tali ipotesi è possibile trovare un intorno $B(x^*, \rho)$ di x^* tale che, per ogni $x \in B(x^*, \rho)$ si abbia:

$$\begin{aligned} h_i(x) &= h_i(x^*) + \nabla h_i(x^*)^T (x - x^*) \quad i = 1, \ldots p \\ g_i(x) &\leqslant g_i(x^*) + \nabla g_i(x^*)^T (x - x^*) \quad i \in I(x^*) \end{aligned}$$

Un caso particolare è quando tutti i vincoli attivia in x^* siano lineari. □

c) **Condizioni di qualificazione dei vincoli di Mangasarian-Fromovitz**

Sia $x^* \in S$ e supponiamo che g, h siano continuamente differenziabili in un intorno di x^*. Si dice che è soddisfatta in x^* la condizione di qualificazione dei vincoli di **Mangasarian-Fromovitz** se:

i. I gradienti dei vincoli di eguaglianza $\{\nabla h_j(x^*), j = 1, \dots, p\}$ sono linearmente indipendenti;

ii. $\nabla g_i(x^*)^T d < 0 \quad \forall i \in I(x^*) \quad \nabla h_i(x^*)^T d = 0 \quad \forall i = 1, \dots, p$;

d) **Condizioni di Slater**

Nelle **condizioni di Slater** consideriamo il caso in cui l'insieme ammissibile è definito attraverso vincoli convessi di diseguaglianza. Quindi:

Supponiamo che le funzioni g_i siano convesso e continuamente differenziabili su un insieme aperto convesso contenente l'insieme ammissibile:

$$S = \{x \in \mathbb{R}^n : g(x) \leqslant 0\}$$

Si dice che è la condizione di Slater è soddisfatta se esiste $\hat{x} \in S$ tale che:

$$g(\hat{x}) < 0$$

Cioè che il punto $\hat{x}$ sia interno all'insieme.

Se vale una di queste condizioni di qualificazione dei vincoli, allora $\lambda_0^* > 0$ e il punto è un **punto regolare**.

Importante 63.

Se $\lambda_0^* > 0$, si può asusmere $\lambda_0^* = 1$ poiché equivale a $(\lambda_0, \lambda, \mu) \to \left(1, \frac{\lambda}{\lambda_0}, \frac{\mu}{\lambda_0}\right)$, lasciando invariato il problema. Inoltre, la Lagrangiana generalizzata diventa **Lagrangiana**:

$$L(x, \lambda, \mu) = 0 \;\; = \;\; f(x) + \sum_{i=1}^{m} \lambda_i \, g_i(x) + \sum_{j=1}^{p} \mu_j \, h_j(x)$$

9.2.2 Condizioni necessarie di KKT (Karush-Kuhn-Tucker)

Sia $x^* \in S$ un punto di minimo locale e supponiamo che f, g, h siano continuamente differenziabili in un intorno di x^*. Supponiamo inoltre che ne punto x^* sia soddisfatta una delle condizioni di qualificazione dei vincoli. Allora esistono $\lambda^* \in \mathbb{R}^m$ e $\mu^* \in \mathbb{R}^p$ tale che:

$$\nabla f(x^*) + \nabla g(x^*) \, \lambda^* + \nabla h(x^*) \mu^* = 0$$

$$g(x^*) \leqslant 0, h(x^*) = 0$$

$${\lambda^*}^T g(x^*) = 0$$

$$\lambda^* \geqslant 0$$

9.2.3 Moltiplicatori di Lagrange

Teorema 64. ***Teorema di Lagrange***

Sia $x^* \in S$ un punto di minimo locale del problema:

$$\min f(x)$$

$$h(x) = 0$$

e supponiamo che f, h siano continuamente differenziabili in un intorno di x^*. Supponiamo inoltre che nel punto x^* sia soddisfatta una delle condizioni di qualificazione dei vincoli.

Allora esiste $\mu^* \in \mathbb{R}^p$ tale che:

$$\nabla f(x^*) + \nabla h(x^*)\, \mu^* = 0$$

$$h(x^*) = 0$$

Se vale il teorema, la ricerca dei punti che soddisfano le condizioni necessarie si riconduce alla ricerca delle soluzioni del sistema di $n+p$ equazioni nelle $n+p$ incogline (x, μ).

9.3 Problemi con vincoli di box

$$\begin{cases} \min f(x) \\ l_i \leqslant x_i \leqslant u_i \end{cases} \quad i = 1, \ldots, n$$

$$l_i - x_i \leqslant 0 \quad i = 1, \ldots, n$$

$$x_i - u_i \leqslant 0 \quad i = 1, \ldots, n$$

Definiamo, inoltre, i moltiplicatori $\hat{\lambda_i}$ e λ_i.

Essendo vincoli lineari, possiamo applicare le condizioni di KKT per ottenere condizioni necessarie di minimo locale. A tale scopo definiamo la Lagrangiana:

$$L(x, \hat{\lambda}, \lambda) = f(x) + \sum_{i=1}^{m} \hat{\lambda_i}\,(l_i - x_i) + \sum_{j=1}^{n} \lambda_i (x_i - u_i)$$

Scriviamo le condizioni di KKT:

(a)

$$\exists (\hat{\lambda}^\star, \lambda^*)\, t.c. \nabla L(x^*, \hat{\lambda}^\star, \lambda^*) = 0 = \nabla f(x^*) - \hat{\lambda}^\star + \lambda^*$$

$$\frac{\partial f(x^*)}{\partial x_i} - \hat{\lambda_i}^\star + \lambda_i^\star = 0 \quad i = 1, \ldots, n$$

(b)

$$\hat{\lambda}_i^*\,(l_i - x_i^*) = 0 \quad i = 1, \ldots n$$

$$\lambda_i^* (x_i^* - u_i) = 0 \quad i = 1, \ldots, n$$

(c)

$$\hat{\lambda}^\star, \lambda_i^* \geqslant 0 \quad i = 1, \ldots, n$$

(d)

$$l_i \leqslant x_i^* \leqslant u_i$$

La condizione (d) si divide in:

$$\begin{cases} x_i^* = l_i \textbf{(1)} \\ l_i < x_i^* < u_i \textbf{(2)} \\ x_i^* = u_i \textbf{(3)} \end{cases}$$

(1) $x_i^* = l_i$

Se sostiuiamo (1) in (b) si ottiene che: λ_i^*=0. Inoltre in (a):

$$\frac{\partial f(x^*)}{\partial x} - \hat{\lambda_i}^\star + \lambda_i^\star = 0 \quad i = 1, \ldots, n$$

Si ottiene:

$$\frac{\partial f(x^*)}{\partial x} = \hat{\lambda_i}^\star \geqslant 0$$

(2) $l_i < x_i^* < u_i$

Se sostituiamo (2) in (b):

$$\hat{\lambda_i}^* (l_i - x_i^*) = 0 \quad i = 1, \ldots n$$

$$\lambda_i^\star (x_i^* - u_i) = 0 \quad i = 1, \ldots, n$$

$(l_i - x_i^*) < 0 \longrightarrow \hat{\lambda_i}^* = 0$

$(x_i^* - u_i) < 0 \longrightarrow \lambda_i^\star = 0$

Inoltre in (a):

$$\frac{\partial f(x^*)}{\partial x} - \hat{\lambda_i}^\star + \lambda_i^\star = 0 \quad i = 1, \ldots, n$$

$$\frac{\partial f(x^*)}{\partial x} = 0$$

(3) $x_i^* = u_i$

Se sostiuiamo (3) in (b) si ottiene che: λ_i^*=0. Inoltre in (a):

$$\frac{\partial f(x^*)}{\partial x} - \hat{\lambda_i}^\star + \lambda_i^\star = 0 \quad i = 1, \ldots, n$$

Si ottiene:

$$\frac{\partial f(x^*)}{\partial x} = \hat{\lambda_i}^\star \leqslant 0$$

$$\frac{\partial f(x^*)}{\partial x} \leqslant 0$$

Ricapitolando siamo tornati al risultato precedente 8.3:

$$\frac{\partial f}{\partial x_i} = \begin{cases} \geqslant 0 & \text{se } x_i^* = l_i \\ = 0 & \text{se } l_i < x_i^* < u_i \\ \leqslant 0 & \text{se } x_i^* = u_i \end{cases}$$

9.4 Condizioni Sufficienti KKT

Le **condizioni di KKT** divengono *condizoni sufficienti (e necessarie se il punto p regolare) di minimo globale* nell'ipotesi che f e g, siano funzioni convesse e che i vincoli di eguaglianza siano lineari.

Teorema 65. *Condizioni sufficienti(e necessarie) di KKT*

Supponiamo che la funzione obiettivo f sia convessa, che le funzioni g_i, per $i = 1, \dots, m$ siano convesse e che i vincoli di uguaglianza siano lineari, ossia: $h(x) = A\,x - b$. Supponiamo inoltre che f,g,h siano continuamente differenziabili su un insieme aperto contenente l'insieme ammissibile. Allora, se esistono moltiplicatori λ^* e μ^* tali che valogno le seguenti condizioni:

$$\begin{aligned} \nabla f(x^*) + \nabla g(x^*)\,\lambda^* + \nabla h(x^*)\,\mu^* = 0 \\ g(x^*) \leqslant 0, h(x^*) = 0 \\ \lambda^{*T} g(x^*) = 0 \\ \lambda^* \geqslant 0 \end{aligned}$$

il punto x^* è un punto di minimo globale vincolato. Se inoltre f è strettamente convessa, allora x^* è l'unico punto di minimo globale vincolato.

Dimostrazione.

Osserviamo inanzitutto che l'insieme ammissibile è convesso e che f eg sono convesse e appartengono a C^1.

Considerato un qualunque punto x ammissibile, essendo $\lambda^* \geqslant 0$, si ha:

$$\begin{aligned} f(x) &\geqslant f(x) + \lambda^{*T} g(x) + \mu^{*T} h(x) \\ h(x) &= h(x^*) + \nabla h(x^*)^T (x - x^*) \\ g_i(x) &\geqslant g_i(x^*) + \nabla g_i(x^*)^T (x - x^*) \end{aligned}$$

Da cui, essendo $\lambda^* \geqslant 0$:

$$\lambda^{*T} g_i(x) \geqslant \lambda^{*T} g_i(x^*) + \lambda^{*T} \nabla g_i(x^*)^T (x - x^*)$$

per la convessità di f:

$$f(x) \geqslant f(x^*) + \nabla f(x^*)^T (x - x^*)$$

Per dimostrare che x^* è minimo globale occorre dimostrare che $f(x^*) \leqslant f(x) \quad \forall x$ ammissibile tale che $g_i(x) \leqslant 0, h_j(x) = 0$

$$\begin{aligned} f(x) &\geqslant f(x) + \lambda^{*T} g(x) + \mu^{*T} h(x) \\ f(x) &\geqslant f(x^*) + \nabla f(x^*)^T (x - x^*) + \lambda^{*T} g_i(x^*) + \lambda^{*T} \nabla g_i(x^*)^T (x - x^*) + \mu^{*T} h(x^*) + \mu^{*T} \nabla h(x^*)^T (x - x^*) \\ &\geqslant f(x^*) + (\nabla f(x^*) + \lambda^{*T} \nabla g_i(x^*)^T + \nabla h(x^*) \mu^*)(x - x^*) = f(x^*) \end{aligned}$$

Quindi abbiamo dimostrato che:

$$f(x) \geqslant f(x^*) \quad \forall x \text{ ammissibile}$$

□

L'iter da seguire è il seguente:

- KKT se soddisfatta allora x^* è un candidato all'ottimo. Altrimenti:
  - x^* è regolare:
    - Si, il punto viene scartato;

– Verificao le condizioni di Fritz John con $\lambda_0 = 0$:

  - Si, allora è un candidato;
  - No, si scara il punto

10 Cenni sulla Teoria Statistica

Consideriamo un problema di **classificazione binario supervisionato** in cui, noto un data set di cui conosco le etichette, vogliamo trarne dei pattern *"nascosti"* all'interno dei dati. Quindi:

$$T = \{(x^i, y^i), x^i \in \mathbb{R}^n, y^i \in \in \{-1, 1\}, i = 1, \dots, l\}$$

Ipotesi:

$\exists$ una distribuzione di probabilità $P(x, y)$ **non nota** che lega x e y.

Inoltre, le osservazionid del training set siano state generate in maniera indipendente e identicamente distribuita secondo $P(x, y)$

Tesi:

Vogliamo implementare una macchina di apprendimento che impari la relazione $x \to y$. Ciò equivale ad ottenre un modello $f(\alpha)\colon \mathbb{R}^n \to \{-1, 1\}$ con $\alpha :=$ **parametro modificabile**.

La $f(\alpha)$ è **deterministica** poiché, una volta fissato α, ad ogni ingresso x fornisce un output y. Quello che dobbiamo fare è addrestare la macchina di apprendimento scegliendo il parametro α.

A tale scopo occorre introdurre una misura che mi quantifica l'errore effettuato nella classificazione del test e minimizzarla. Questa misura viene detta **rischio effettivo**(coincide con il valore atteso $\mathbb{E}$ data la probabilità):

$$R(\alpha) = \int \frac{1}{2} |y - f(x, \alpha)| \mathrm{d}P(x, y)$$

In cui:

- y è l'uscita reale;
- $f(x, \alpha)$ è l'uscita ottenuta dalla macchina;

Tuttavia il **rischio effettivo** non è calcolabile dato che abbiamo supposto che la distribuzione di probabilità non è nota. Un'alternativa viene data dal **rischio empirico:**

$$R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} |y^i - f(x^i, \alpha)|$$

Occorre notare che abbiamo scorrelato il rischio dalla conoscenza della distribuzione di probabilità, ma l'abbiamo legata ai dati del nostro training set e quindi è una misura che possiamo calcolare.

Nota 66.

In generale non vi è una relazione tra rischio empirico e rischio effettivo.

Inoltre, se confrontassimo due rischi empirici di due macchine diverse anche se possono avere lo stesso valore si potrebbero comportare in maniera diversa in presenta di dati del test set.

10.1 Teoria di Vapinil-Chervonenkis

Sia $\eta \in [0,1]$ con probabilità $1-\eta$ vale la seguente diseguaglianza:

$$R(\alpha) \leqslant R_{\text{emp}} + \sqrt{\frac{h \log\left(\frac{2\,l}{h}\right) - \log\left(\frac{\eta}{4}\right)}{l}}$$

In cui definiamo la **VC Confidence**:

$$\text{VCC} = \sqrt{\frac{h \log\left(\frac{2\,l}{h}\right) - \log\left(\frac{\eta}{4}\right)}{l}}$$

essa dipende da η, l, h. L'ultimo parametro è detto **VC Dimension** e misura la capacità di classificazione di una macchina $f(\alpha)$.

Quindi il termine a sinistra è composto da una media pesata degli errori del training set più la VC Confidenze che dipende dalla complessità della macchina, dal numero di feauture e da η, costituendo così un upper-bound per l'errore reale.

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \text{VC}(h, \eta, l) = U(\alpha, h)$$

Nella progettazione della nostra macchina vogliamo un **trade-off** fra complessità della macchina e il rischio reale. A tale scopo, occorre minimizzare l'upper bound $U(\alpha, h)$.

Definizione 67. *VC dimension*

*h è il massimo numero di punti che possono essere **frammentati**, ovvero che possono essere separati tramite la macchina di apprendimento $f(\alpha)$ in due classi quando etichettati in tutti i modi possibili.*

La nuova funzione obiettivo da massimizzare è il **rischio strutturale**:

$$U(\alpha, h) = R_{\text{emp}}(\alpha) + \text{VCC}(h, \alpha)$$

Si definiscono delle classi di funzioni di apprendimento:

$$F_1 \subseteq F_2 \subseteq \cdots \subseteq F_p$$

$$h_1 \leqslant h_2 \leqslant \cdots \leqslant h_p$$

Procedura euristica per minimizzare il rischio strutturale

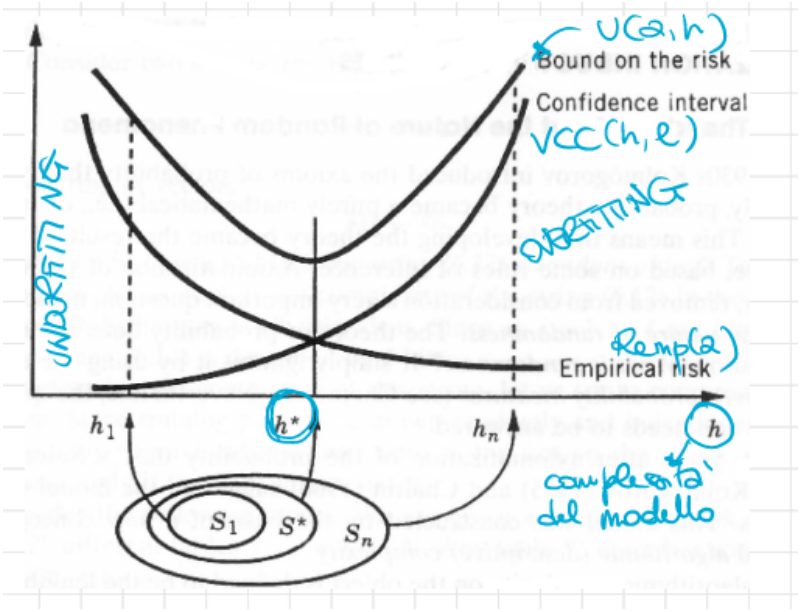
Per ogni classi F_j con VC dimension h_j, minimizza il rischio empirico $R_{\text{emp}}(\alpha)$:

$$\alpha_j = \arg\min_{F_j} R_{\text{emp}}(\alpha)$$

Inoltre, calcola:

$$U^j = R_{\text{emp}}(\alpha_j) + \text{VCC}(h_j, l) \quad j = 1, \ldots p$$

Scelgo la classe di macchine ch emi restituisce il valore più basso di U^j.



Nota 68. h non è necessariamente proporzionale alla dimensione di α.

Si hanno due approcci possibili per minimizzare $U(\alpha, w)$:

1. Fisso h e minimizzo $R_{\text{emp}}(\alpha) \longrightarrow$ **reti neurali**;
2. Fisso $R_{\text{emp}}(\alpha) = 0$ e minimizzo $\text{VCC}(h, l) \longrightarrow$ **SVM** (Super Vector Machine);

11 Iperpiani Orientati

Un **iperpiano orientato** è definito come:

$$f(\alpha) = \text{sgn}(w^T x + b) \quad \alpha = \begin{pmatrix} w \\ b \end{pmatrix} \quad x \in \text{TS}$$

Teorema 69.

Un insieme di punti $\{x^1, \ldots, x^m\}$ *è frammentabile dalla famiglia degli iperpiani orientati* $f(\alpha) = \text{sgm}(w^T x + b) \iff \{x^1, \ldots, x^m\}$ *sono* ***affinemente indipendenti*** *cioè* $x_2 - x_1, \ldots, x_m - x_1$ *sono linearmente indipendenti.*

Importante 70.

h è il massimo numero di punti affinemente indipendenti in $\mathbb{R}^n$, quindi $h = n + 1$.

Definizione 71. *Margine*

Il ***margine*** *è la distanza minimi dell'iperpiano orientato dai punti del training set.*

Intuizione

Quanto più l'iperpiano è lontano dai dati di training, tanto più sarà migliore la sua capacità di generalizzazione.

Idea

Ci si restringe ad un sottocaso degli iperpiani detti **iperpiani orientati con gap di tolleranza**(margine).

Ipotesi: Supponiamo che i miei dati appartengano ad una sgera di diametro D.

$$y = H(w,b) = \begin{cases} 1 & \text{se } \frac{w^T x + b}{\|w\|} \geqslant \rho \\ -1 & \text{se } \frac{w^T x + b}{\|w\|} \leqslant -\rho \end{cases}$$

Nota 72.

- $\frac{w^T x + b}{\|w\|} :=$ è la distanza di x dall'iperpiano;
- $\rho = 0 \longrightarrow$ iperpiani orientati.

Ora devo calcolare la VC dimensione, essa dipende sia da ρ che da D.

11.1 Iperpiani con gap di tolleranza

Supponiamo che i dati siano generati all'interno di una sfera di diametro D e che si assegnino le uscite:

$$y = \begin{cases} 1 & \text{se } \frac{w^T x + b}{\|w\|} \geqslant \rho \\ -1 & \text{se } \frac{w^T x + b}{\|w\|} \leqslant -\rho \end{cases}$$

Definizione 73. *Insiemi del data set*

$$A = \{x^i \in T : y^i = 1\}$$

$$B = \{x^j \in T : y^j = -1\}$$

11.1.1 Variazione di VC dimension

L'intuizione è che all'aumentare di ρ, h diminuisce.

Teorema 74. *VC dimension*

Per gli iperpiani con margine si ha:

$$h \leqslant \min\left\{\left\lceil \frac{D^2}{\rho^2} \right\rceil, n\right\} + 1$$

Nella minimzzazione del *rischio strutturale* si prende:

$$F_i = \{f(x) = w^T x + b, \rho(w,b) \geqslant \rho_i\}$$

Definizione 75. *Margine Iperpiano con gap di tolleranza*

Definiamo il ***margine dell'iperpiano*** *come:*

$$\rho(w,b) = \min_{x^i \in T} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

Data una famglia di iperpiani F_i vogliamo determinare l'iperpiano che minimizza l'upper-bound così definito:

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \text{VCC}(h, l, \eta) = U(\alpha, h)$$

così facendo otteniamo l'**iperpiano ottimo** che ha un margine massimo ed ha un livello di generalizzazione maggiore su nuovi dati appartenenti al test set.

Ipotesi: A, B sono lineamente separabili e quindi:

$$\exists(\overline{w}, \bar{b}) \text{ che speara correttamente } T \text{ in due insemi } A e B$$

In particolare, dato un *training set*:

$$\mathrm{TS} = \{(x^i, y^i)\, i = 1, \ldots, l, x^i \in \mathbb{R}^n . y^i \in \{-1, 1\}\}$$

$$A = \{x^i \in T \colon y^i = 1\}$$

$$B = \{x^j \in T \colon y^j = -1\}$$

I due insiemi sono **linearmente separabili** se $\exists(w, b) \colon y(x) = \mathrm{sgn}(w^T x + b)$ classifica correttamente il TS, ovvero:

$$w^T x^i + b > 0 \qquad x^i \in A$$

$$w^T x^j + b < 0 \qquad x^j \in B$$

Poiché i risolutori si comportano male in presenza di disequazioni non strette, riscriviamo le relazioni superiori nel seguente modo:

$$w^T x^i + b \geqslant \varepsilon \qquad x^i \in A$$

$$w^T x^j + b \leqslant -\varepsilon \qquad x^j \in B$$

Inotre, dato che ε è uno scalare >0, possiamo rinominare $w = \frac{w}{\varepsilon}$ e $b = \frac{b}{\varepsilon}$ ottenendo:

$$w^T x^i + b \geqslant 1 \qquad x^i \in A$$

$$w^T x^j + b \leqslant -1 \qquad x^j \in B$$

Che corrisponde ad avere un $R_{\mathrm{emp}}(w, b) = 0$.

Fra tutti questi iperpiano, come da ipotesi, vofliamo determinare quello ottimo e con il margine massimo:

$$\begin{aligned} \max_{w,b} \rho(w, b) &= \min_{x^i \in A \cup B} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\} \\ w^T x^i + b &\geqslant 1 \quad \forall x^i \in A \\ w^T x^j + b &\leqslant -1 \quad \forall x^j \in B \end{aligned}$$

Questo problema combina un problema di massimo e di minimo, il che può essere riscritto come:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ w^T x^i + b \quad & \geqslant \quad 1 \quad \forall x^i \in A \\ w^T x^j + b \quad & \leqslant \quad -1 \quad \forall x^j \in B \end{aligned}$$

Nota 76.

Questo problema è convesso poiché, i vincoli sono lineare e la funzione obiettivo è quadratica e anch'essa convessa.

Lemma 77.

Per ogni iperpiano ammissibile $(\hat{w}, \hat{b})$ *si ha:*

$$\rho(\hat{w}, \hat{b}) \geqslant \frac{1}{\|\hat{w}\|}$$

Dimostrazione.

$$\rho(w, b) \;=\; \min_{x^i \in A \cup B} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

Poiché è ammissibile:

$$w^T x^i + b \;\geqslant\; 1 \quad \forall x^i \in A \;\rightarrow \forall x^i \in A \quad |w^T x^i + b| \geqslant 1 = w^T \hat{x}^i + \hat{b}$$

$$w^T x^j + b \;\leqslant\; -1 \quad \forall x^j \in B \;\rightarrow |w^T x^i + b| = -(w^T x^i + b) \geqslant 1$$

Quindi:

$$|w^T x^i + b| \geqslant 1 \quad \forall x^i \in A \cup B$$

$$\min_{x^i \in A \cup B} \{|w^T x^i + b|\} \geqslant 1 \longrightarrow \rho(\hat{w}, \hat{b}) \geqslant \frac{1}{\|\hat{w}\|} \qquad \square$$

Teorema 78.

Per ogni iperpiano ammissibile $(\hat{w}, \hat{b})$*, essite un altro iperpiano ammissibile* $(\overline{w}, \bar{b})$ *tale che:*

$$\rho(\overline{w}, \bar{b}) = \frac{1}{\|w\|} \geqslant \rho(\hat{w}, \hat{b})$$

Inoltre, esistono due punti $x^+ \in A$ *e* $x^- \in B$ *tali che:*

$$w^T x^+ + b = 1 \qquad\qquad w^T x^- + b = -1$$

Dimostrazione.

Prendo $x^i \in A$ e $x^j \in B$ più vicini, per calcolare la distanza definisco:

$$\hat{d}_i = \min_{x^i \in A} \left\{ \frac{|\hat{w}^T x^i + \hat{b}|}{\|\hat{w}\|} \right\}$$

$$\hat{d}_j = \min_{x^j \in B} \left\{ \frac{|\hat{w}^T x^j + \hat{b}|}{\|\hat{w}\|} \right\}$$

Per l'insieme A:

$$\hat{d}_i = \frac{|\hat{w}^T \hat{x}^i + \hat{b}|}{\|\hat{w}\|} \leqslant \frac{|\hat{w}^T x^i + \hat{b}|}{\|\hat{w}\|} \quad \forall x^i \in A \longrightarrow \hat{w}^T \hat{x}^i \leqslant \hat{w}^T x^i$$

Per l'insieme B:

$$\hat{d}_j = \frac{|\hat{w}^T \hat{x}^j + \hat{b}|}{\|\hat{w}\|} \leqslant \frac{|\hat{w}^T x^j + \hat{b}|}{\|\hat{w}\|} \quad \forall x^j \in B$$

$$\frac{-\hat{w}^T \hat{x}^j - \hat{b}}{\|\hat{w}\|} \leqslant \frac{-\hat{w}^T x^j - \hat{b}}{\|\hat{w}\|} \longrightarrow -\hat{w}^T \hat{x}^j \leqslant -\hat{w}^T x^j$$

Definiamo $\rho(\hat{w}, \hat{b}) = \min\{\hat{d}_i, \hat{d}_j\}$

Corollario 79. *Proprietà del minimo*

$$\min\{a,b\} \leqslant \frac{1}{2}a + \frac{1}{2}b$$

Dimostrazione.

$$\min\{a,b\} \leqslant a \qquad \min\{a,b\} \leqslant b \longrightarrow \frac{1}{2}\min\{a,b\} \leqslant \frac{1}{2}a \qquad \frac{1}{2}\min\{a,b\} \leqslant \frac{1}{2}b$$

Sommo entrambe le relazioni:

$$\min\{a,b\} \leqslant \frac{1}{2}a + \frac{1}{2}b \qquad \square$$

Quindi, ritornando alla prima dimostrazione:

$$\begin{aligned}\rho(\hat{w},\hat{b}) = \min\{\hat{d}_i,\hat{d}_j\} \leqslant \frac{1}{2}\hat{d}_i + \frac{1}{2}\hat{d}_j &= \frac{1}{2}\frac{|\hat{w}^T\hat{x}^i + \hat{b}|}{\|\hat{w}\|} + \frac{1}{2}\frac{|\hat{w}^T\hat{x}^j + \hat{b}|}{\|\hat{w}\|} \\ &= \frac{1}{2}\frac{\hat{w}^T\hat{x}^i + \hat{b}}{\|\hat{w}\|} - \frac{1}{2}\frac{\hat{w}^T\hat{x}^j + \hat{b}}{\|\hat{w}\|} \\ &= \frac{\hat{w}^T(\hat{x}^i - \hat{x}^j)}{2||w|}\end{aligned}$$

Modifico $(\hat{w},\hat{b})$ introducendo due scalari α, coefficiente di rotazione, e β, coefficiente di traslazione, $\longrightarrow(\overline{w},\bar{b})$:

$$\overline{w} = \alpha\,\hat{w}$$

$$\bar{b} = \beta$$

Questi due parametri devono soddisfare le seguenti relazioni:

$$\begin{cases} \alpha\,\hat{w}^T\hat{x}^i + \beta = 1 \\ \alpha\hat{w}^T\hat{x}^j + \beta = -1 \end{cases}$$

Che rappresentano i punti sul margine dell'iperpiano. Equivalgono a vincoli attivi cioè soddisfatti all'uguaglianza. Risolviamo ora il sistema:

$$\alpha\,\hat{w}^T(\hat{x}^i - \hat{x}^j) = 2 \rightarrow \alpha = \frac{2}{\hat{w}^T(\hat{x}^i - \hat{x}^j)} \geqslant 2$$

$$\beta = 1 - \alpha\hat{w}^T\hat{x}^i = 1 - \frac{2\hat{w}^T\hat{x}^i}{\hat{w}^T(\hat{x}^i - \hat{x}^j)} = \frac{\hat{w}^T\hat{x}^i - \hat{w}^T\hat{x}^j - 2\,\hat{w}^T\hat{x}^i}{\hat{w}^T(\hat{x}^i - \hat{x}^j)} = \frac{\hat{w}^T(\hat{x}^i + \hat{x}^j)}{\hat{w}^T(\hat{x}^i - \hat{x}^j)}$$

Nota 80.

$$0 < \alpha < \leqslant 1$$

A questo punto dobbiamo verificare che il nuovo iperpiano $(\overline{w},\bar{b})$ sia ammissibile. A tale scopo, so che:

$$1 \leqslant \hat{w}^T\hat{x}^i + \hat{b} \leqslant \hat{w}^T\hat{x}^j + \hat{b} \quad \forall x^i \in A$$

$$\begin{aligned} |\hat{w}^T\hat{x}^i + \hat{b}| &\leqslant |\hat{w}^Tx^i + \hat{b}| \qquad \forall x^i \in A \\ |\hat{w}^T\hat{x}^j + \hat{b}| &\leqslant |\hat{w}^Tx^j + \hat{b}| \qquad \forall x^j \in B \longrightarrow |\hat{w}^T\hat{x}^j + \hat{b}| \geqslant 1 \end{aligned}$$

Quindi:

$$\begin{aligned} \overline{w}^T x^i + \bar{b} &= \alpha\, \hat{w}^T x^i + \beta \quad \text{Ricordando che } \hat{w}^T \hat{x}^i \leqslant \hat{w}^T x^i \\ &\geqslant \alpha\, \hat{w}^T \hat{x}^i + \beta = 1 \quad \forall x^i \in A \end{aligned}$$

e

$$\begin{aligned} \overline{w}^T x^j + \bar{b} &= \alpha\, \hat{w}^T x^j + \beta \quad \text{Ricordando che } -\hat{w}^T \hat{x}^j \leqslant -\hat{w}^T x^j \\ &\leqslant \alpha\, \hat{w}^T \hat{x}^j + \beta = -1 \quad \forall x^j \in B \end{aligned}$$

Quindi l'iperpiano ottenuto ruotando e traslando è ancora ammissibile.

In conclusione, $(\overline{w}, \bar{b})$ è l'iperpiano di separazione:

$$\rho(\overline{w}, \bar{b}) = \min_{x^i \in A \cup B} \left\{ \frac{|\overline{w}^T x^i + \hat{b}|}{\|\overline{w}\|} \right\} = \frac{1}{\|\overline{w}\|} = \frac{1}{\alpha \|\hat{w}\|}$$

Nota 81.

Il numeratore vale 1 per costruzione.

$$= \frac{1}{\alpha \|\hat{w}\|} = \frac{\hat{w}^T (\hat{x}^i - \hat{x}^j)}{2\|\hat{w}\|} = \frac{1}{2} \hat{d}_i + \frac{1}{2} \hat{d}_j \geqslant \rho(\hat{w}, \hat{b})$$

Inoltre, so che:

$$\begin{cases} \overline{w}^T \hat{x}^i + \beta = 1 \\ \overline{w}^T \hat{x}^j + \beta = -1 \end{cases} \longrightarrow \begin{cases} x^+ = x^i \\ x^- = x^j \end{cases}$$

Che erano l'ipotesi del teorema. □

Ricapitolando, l'**iperpiano ottimo:**

$$\max_{w, \rho} \rho(w, b) = \min_{x^i \in A \cup B} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

$$w^T x^i + b \geqslant 1 \quad \forall x^\iota \in A$$

$$w^T x^j + b \leqslant -1 \quad \forall x^j \in B$$

Lemma 82.

Per ogni iperpiano ammissibile $(\hat{w}, \hat{b})$ *si ha* $\rho(\hat{w}, \hat{b}) \geqslant \frac{1}{\|\hat{w}\|}$

Teorema 83.

Per ogni iperpiano ammissibile $(\hat{w}, \hat{b})$ *esiste un iperpiano* $(\overline{w}, \bar{b})$ *tale che:*

$$\rho(\overline{w}, \bar{b}) = \frac{1}{\|\overline{w}\|} \geqslant \rho(\hat{w}, \hat{b})$$

Inoltre esistono $x^+ \in A, x^- \in B$ *tali che:*

$$\overline{w} x^+ + \bar{b} = 1$$

$$\overline{w} x^- + \bar{b} = -1$$

Vogliamo far vedere che il problema inizialia di max min è equivalente a :

$$\min \frac{1}{2}\|w\|^2 = f(w,b)$$

$$w^T x^i + b \geqslant 1 \quad \forall x^i \in A$$

$$w^T x^j + b \leqslant -1 \quad \forall x^j \in b$$

L'insieme ammissibile è chiuso ma potrebbe essere illimitato. Inoltre, è anche convesso e non vuoto:

$$\nabla f = \begin{pmatrix} w \\ 0 \end{pmatrix} \qquad \nabla^2 f = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \succcurlyeq 0$$

Quindi f è convessa e coerciva solo rispetto a w. Tuttavia, non è ne convessa ne coerciva ne vale Weierstrass-Ermann, occorre quindi determinare l'unicità della soluzione.

Teorema 84.

Il problema equivalente ammette una soluzione unica (w^*, b^*).

Dimostrazione.

L'insieme ammissibile è chiuso, ma non limitato. Dobbiamo, quindi, far vedere l'insieme di livello è compatto, se lo è allora il problema ammette soluzione.

$$L_0 = \left\{ (w,b) \in F : \frac{1}{2}\|w\|^2 \leqslant \frac{1}{2}\|w_0\|^2 \right\}$$

$$F(w,b) : w^T x^i + b \geqslant 1 \quad \forall x^i \in A, w^T x^j + b \leqslant -1 \quad \forall x^j \in b$$

L_0 è chiuso poiché F è chiuso, devo far vedere che è limitato.

Suppongo per assurdo che $\exists \{w^k, b^k\}$ tale che $w^k, b^k \in L_0 \quad \left\| \begin{pmatrix} w^k \\ b^k \end{pmatrix} \right\| \uparrow \infty$

Dato che l'insieme di livello è definito come $\|w^k\|^2 \leqslant \|w_0\|^2 \longrightarrow \|w^k\|$ è finita, suppongo che $\lim_{b \to \infty} b^k = \infty$:

$$\begin{pmatrix} w^k \\ b^k \end{pmatrix} \in F \longrightarrow \begin{array}{l} {w^k}^T x^i + b^k \geqslant 1 \quad \forall x^i \in A \\ {w^k}^T x^j + b^k \leqslant -1 \quad \forall x^j \in B \end{array}$$

Se $b^k \to \infty$ il primo gruppo di vincoli è soddisfatto. Ci rimane da soddisfare il secondo:

$${w^k}^T x^j + b^k \leqslant -1 \quad \forall x^j \in B$$

Per essere soddisfatto deve valere ${w^k}^T x^j \to -\infty \longrightarrow \|w^k\| \to \infty$. Ciò comporta un assurdo poiché la successione non appartiene più a L_0.

Allo stesso modo se $b^k \to -\infty$ devo soddisfare:

$$w^{k^T} x^i + b^k \geqslant 1 \quad \forall x^i \in A \longrightarrow w^{k^T} x^i \to \infty \longrightarrow \|w^k\| \to \infty$$

Ciò comporta un assurdo. Quindi, L_0 è un insieme compatto e quindi esiste soluzione.

L'unico aspetto che ci occorre dimostrare che la soluzione è unica: suppongo per assurdo che non lo sia. In particolare:

$$\exists (w^*, b^*), (\overline{w}, \bar{b}) \text{ ottime}$$

Quindi:

$$\frac{1}{2}\|w^*\|^2 = \frac{1}{2}\|\overline{w}\|^2 \leqslant \frac{1}{2}\|w\|^2 \quad \forall \begin{pmatrix} w \\ b \end{pmatrix} \text{ammissibile}$$

Suppongo che $\overline{w} \neq w^*$. Siccome l'insieme è convesso, tramite la definizione di convessità:

$$\exists \lambda \in [0,1] \quad \lambda \begin{pmatrix} w^* \\ b^* \end{pmatrix} + (1-\lambda) \begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix} \in F$$

Scegliamo $\lambda = \frac{1}{2}$:

$$\begin{pmatrix} \tilde{w} \\ \tilde{b} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} w^* \\ b^* \end{pmatrix}$$
$$f(\tilde{w}, \tilde{b}) = \frac{1}{2} \left\| \frac{1}{2}\overline{w} + \frac{1}{2} w^* \right\|^2$$

Calcoliamo il valore della funzione obiettivo. La funzione obiettivo è stressamente convessa in w se e solo se:

$$f\left(\lambda \begin{pmatrix} w \\ b \end{pmatrix} + (1-\lambda) \begin{pmatrix} w^* \\ b^* \end{pmatrix} \right) < \lambda f \begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix} + (1-\lambda) \begin{pmatrix} w^* \\ b^* \end{pmatrix}$$

In $\tilde{w}$:

$$f(\tilde{w}, \tilde{b}) = \frac{1}{2}\|\tilde{w}\|^2 < \frac{1}{2}\frac{1}{2}\|w^*\| + \frac{1}{2}\frac{1}{2}\|\overline{w}\|^2 = \frac{1}{2}\|w^*\|^2$$

Questo è un assurdo poiché avremmo travo un iperpiano ammissibile migliore di quello ottimo e quindi deve valere necessariamente che:

$$\overline{w} = w^*$$

Ora ci rimane da controllare il valore del termine noto b: supponiamo $b^* > \bar{b}$.

So che $\exists x^+ \in A$:

$$w^{*^T} x^+ + b^* = 1 = \overline{w}^T x^+ + b^* > \overline{w}^T x^+ + \bar{b} \longrightarrow \overline{w}^T x^+ + \bar{b} < 1 \longrightarrow \text{ASSURDO}$$

Questo è un assurdo poiché $x^+ \in A$ e $\begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix}$ è ammissibile. Allo stesso modo, supponiamo che $b^* < \bar{b}$:

$$w^{*^T} x^- + b^* = -1 = \overline{w}^T x^- + b^* < \overline{w}^T x^- + b^* \to \overline{w}^T x^- + \bar{b} > 1 \longrightarrow \text{ASSURDO}$$

Questo è un assurdo poiché $x^- \in B$ e $\begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix}$ è ammissibile.

Quindi per valere entrambe le condizioni, dobbiamo avere che:

$$\bar{b} = b^*$$

La soluzione, infine, è unica. □

Teorema 85.

L'unica soluzione $\begin{pmatrix} w^* \\ b^* \end{pmatrix}$ *del problema:*

$$\min \frac{1}{2}\|w\|^2$$

$$w^T x^i + b \geqslant 1 \quad \forall x^i \in A$$

$$w^T x^j + b \leqslant -1 \quad \forall x^j \in b$$

è l'unica soluzione del problema:

$$\max \rho(w, b)$$

$$w^T x^i + b \geqslant 1 \quad \forall x^i \in A$$

$$w^T x^j + b \leqslant -1 \quad \forall x^j \in B$$

Dimostrazione.

Dal lemma 82 so che per ogni iperpiano ammissibile:

$$\rho(\hat{w}, \hat{b}) \geqslant \frac{1}{\|\hat{w}\|}$$

Inoltre, so che per ogni iperpiano ammissibile $\begin{pmatrix} \hat{w} \\ \hat{b} \end{pmatrix}$ esiste un iperpiano $\begin{pmatrix} \overline{w} \\ \bar{b} \end{pmatrix}$ tale che:

$$\rho(\hat{w}, \hat{b}) \leqslant \rho(\overline{w}, \bar{b}) = \frac{1}{\|\overline{w}\|}$$

In aggiunta, so che:

$$\frac{1}{2}\|w^*\|^2 \leqslant \frac{1}{2}\|w\|^2 \quad \forall \begin{pmatrix} w \\ b \end{pmatrix} \text{ammissibile}$$

$$\longrightarrow \|w^*\| \leqslant \|w\| \quad \forall \begin{pmatrix} w \\ b \end{pmatrix} \text{ammissibile}$$

$$\longrightarrow \frac{1}{\|w^*\|} \geqslant \frac{1}{\|w\|} \quad \forall \begin{pmatrix} w \\ b \end{pmatrix} \text{ammissibile}$$

Per qualunque $\begin{pmatrix} \hat{w} \\ \hat{b} \end{pmatrix}$ ammissibile:

$$\frac{1}{\|\hat{w}\|} \leqslant \rho(\hat{w}, \hat{b}) \leqslant \frac{1}{\|\overline{w}\|} \leqslant \frac{1}{\|w^*\|}$$

Allora:

$$\rho(\hat{w},\hat{b}) \leqslant \rho(w^*,b^*) \quad \forall \begin{pmatrix} \hat{w} \\ \hat{b} \end{pmatrix}$$

Cioè è l'iperpiano a massimo margine.

Abbiamo dimostrato che è l'**ottimo globale.** □

Il problema della riformulazione del problema è dato dalla presenza di vincoli per ogni punto e quindi risulta oneroso dal punto di vista computazionale.

12 Teoria della Dualità

12.1 Introduzione

Il problema che vogliamo risolvere è:

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ y^i(w^T x^i + b) \geqslant 1 \quad \forall i = 1,\dots,l \end{cases}$$

A tale scopo generiamo una formulazione alternativa del problema, detto problema duale, e, di conseguenza, otteniamo i seguenti vantaggi:

1. Teorici: otteniamo risultati e informazioni sul problema originario⟶condizioni di ottimo;

2. Praciti: il duale è più semplice da risolvere⟶ il numero di vincoli diminuisce mentre il numero delle variabili aumentano.

Quindi la situazioni che abbiamo davanti è del tipo:

Problema Primale

$$\begin{cases} \min f(x) \\ x \in S \end{cases}$$

Problema Duale

$$\begin{cases} \max \psi(u) \\ u \in U \end{cases}$$

Tipicamente vale la proprietà di **dualità debole:**

$$\sup \psi(u) \leqslant \inf f(x)$$

$$u \in U \qquad x \in S$$

In alcuni casi, cioè quando il problema è convesso, vale la **dualità forte**:

$$\sup \psi(u) = \inf f(x)$$

$$u \in U \qquad x \in S$$

La dualità ci permette di ottenere:

1. Stima del valore ottimo (lower-bound) del primale;
2. Condizioni di ottimo;
3. Algoritmi di soluzione.

12.2 Dualità di Wolfe

Dato un problema **primale** non lineare convesso, del tipo:

$$\begin{cases} \min f(x) \\ g_i(x) \leqslant 0 \quad i = 1, \dots, m \\ h_j(x) = 0 \quad j = 1, \dots, p \end{cases}$$

$$f, g_i, h_j \in C^1 \quad i = 1, \dots, n \quad j = 1, \dots, p$$

In particolare:

$$f, g_i \quad i = 1, \dots, m \quad \text{convesse}$$

$$h_j \quad j = 1, \dots, p \qquad \text{lineare}$$

Quindi abbiamo che le condizioni di KKT risultano necessarie e sufficienti di ottimo globale:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$$

Definiamo a questo punto il **duale di Wolfe**:

$$\begin{cases} \max_{x,\lambda,\mu} L(x, \lambda, \mu) \\ \nabla_x L(x, \lambda, \mu) = 0 = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) \\ \lambda \geqslant 0 \end{cases}$$

Teorema 86.

*Se il problema **primale** che ammette soluzione x^* con moltiplicatori (λ^*, μ^*) allora (x^*, λ^*, μ^*) è soluzione del duale e il **gap di dualità** è nullo, cioè:*

$$f(x^*) = L(x^*, \lambda^*, \mu^*)$$

Dimostrazione.

(x^*, λ^*, μ^*) è soluzione del problema primale $\longleftrightarrow$le condizioni di KKT sono soddisfatte:

$$\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) = 0 \quad & (1) \\ \lambda^* \geqslant 0 \quad & (2) \\ \lambda_i g_i(x^*) = 0 \qquad & i = 1, \dots, m \\ g_i(x^*) \leqslant 0 \qquad & i = 1, \dots, m \\ h_j(x^*) = 0 \qquad & j = 1, \dots, p \end{aligned}$$

La (1)+(2) forniscono la soluzione (x^*, λ^*, μ^*) ammissibile per i duale.

$$L(x^*, \lambda^*, \mu^*) = f(x^*) + \sum_{i=1}^{m} \lambda_i \, g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) = f(x^*)$$

Nota 87.

- $\sum_{i=1}^{m} \lambda_i \, g_i(x) = 0$ per complementarietà;
- $\sum_{j=1}^{p} \mu_j h_j(x) = 0$ per ammissibilità

Quindi il gap di dualità è nullo.

A questo punto devo far vedere che la soluzione è **ottima** per il duale. Quindi devo far vedere che:

$$L(x^*, \lambda^*, \mu^*) \geqslant L(x, \lambda, \mu) \quad \forall (x, \lambda, \mu) \text{ ammissibile per il duale}$$

$$L(x^*, \lambda^*, \mu^*) = f(x^*) \geqslant f(x^*) + \sum_{i=1}^{m} \lambda_i \, g_i(x^*) + \sum_{j=1}^{p} \mu_j h_j(x^*) = L(x^*, \lambda, \mu) \quad (\alpha)$$

Nota 88.

- λ_i poiché λ è ammissibile per il duale;
- $g_i(x^*) \leqslant 0$ per l'ammissibilità di x^* nel problema primale;
- $h_j(x^*) = 0$ poiché x^* è ammissibile

Otteniamo quindi:

$$L(x^*, \lambda^*, \mu^*) \geqslant L(x^*, \lambda, \mu)$$

Inoltre sappiamo che $L(x, \lambda, \mu)$ è convessa e appartiene a C^1. Allora:

$$L(x^*, \lambda, \mu) \geqslant L(x, \lambda, \mu) + \nabla_x L(x, \lambda, \mu)^T (x^* - x) \quad \forall (x, \lambda, \mu) \text{ ammissibile nel duale } (\beta)$$

Nota 89.

- $\nabla_x L(x, \lambda, \mu)^T (x^* - x) = 0$ per l'ammissibilità duale

Infine, unendo $(\alpha) + (\beta)$, si giunge a:

$$L(x^*, \lambda^*, \mu^*) \geqslant L(x, \lambda, \mu) \quad \forall (x, \lambda, \mu) \text{ ammissibil per il duale}$$

Il che implica che la terna (x^*, λ^*, μ^*) è ottima per il duale. □

Nota 90.

Dopo aver effettuato queste operazioni di trasformazione del problema primale, il duale non è più lineare e non vale neanche la convessità. Ciò corrisponde ad un aumento di complessità. Tuttavia nel nostro caso, con funzione obiettivo quadratica e vincoli lineari, il problema si semplifica.

12.3 Programmazione Quadratica - Duale

Dato il problema dell'iperpiano ottimo:

$$\begin{cases} \min \frac{1}{2} x^T Q\, x + c^T x \\ A\, x \geqslant b \\ A \in \mathbb{R}^{m \text{x} n}, b \in \mathbb{R}^m \end{cases} \quad Q \succcurlyeq 0$$

Definiamo la Lagrangiana.

$$L(x, \lambda) = \frac{1}{2} x^T Q\, x + c^T x + \lambda^T (A\, x - b)$$

Scriviamo il duale:

$$\max_{x,\lambda} \frac{1}{2} x^T Q\, x + c^T x + \lambda^T (A\, x - b)$$

$$Q\, x + c + A^T \lambda = 0$$

$$\lambda \geqslant 0$$

Moltiplicando il vincolo per x^T:

$$x^T Q\, x + c^T x - x^T A^T \lambda = 0 \longrightarrow c^T x + \lambda^T A\, x = -x^T Q\, x$$

e sostituisco nella funzione obiettivo:

$$\max \frac{1}{2} x^T Q\, x - x^T Q\, x - \lambda^T b = -\frac{1}{2} x^T\, Q\, x - \lambda^T b$$

$$\longrightarrow -\min \frac{1}{2} x^T Q\, x + b^T \lambda$$

Quindi posso ora scrivere il duale:

$$\begin{cases} -\min_{x,\lambda} \frac{1}{2}\, x^T Q\, x + b^T \lambda = \theta(x, \lambda) \\ Q\, x + c + A^T \lambda = 0 \\ \lambda \geqslant 0 \end{cases} \quad \textbf{Problema Convesso}$$

$$\nabla^2 \theta(x, \lambda) = \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \succcurlyeq 0 \quad \text{semidefinito positivo} \rightarrow \text{solo strettamente convessa}$$

Quindi le condizioni di KKT sono necessarie e sufficienti di ottimo globale.

Teorema 91.

Dato il problema primale e duale:

$$\begin{cases} \min \frac{1}{2}\, x^T Q\, x + c^T x \\ A\, x \leqslant b \end{cases} (P) \qquad \begin{cases} -\min \frac{1}{2} x^T Q\, x + b^T \lambda \\ Q\, x + c + A^T \lambda = 0 \qquad [v] \\ \lambda \geqslant 0 \qquad\qquad\quad [z] \end{cases} (D)$$

La condizione necessaria è che $Q \succcurlyeq 0$.

Sia $(\overline{x}, \overline{\lambda})$ soluzione del duale, allora $\exists x^*$, non necessariamente uguale a $\overline{x}$, tale che:

1. $Q(\overline{x} - x^*) = 0$;
2. *x^* soluzione del problema primale (P);*
3. *$(\overline{x}, \overline{\lambda})$ è una coppia di minimo globale-moltiplicatore del primale (P);*

Nota 92.

La x, passando dal primale al duale può cambiare, ma la $\overline{\lambda}$ è lo stesso.

Dimostrazione.

$(\overline{x}, \overline{\lambda})$ è soluzione ottima del duale $(D) \longrightarrow$ KKT è soddisfatta nel duale.

Scriviamo ora la Lagrangiana del duale W con le corrispettive condizioni di KKT:

$$
\begin{aligned}
W(x, \lambda, v, z) = \frac{1}{2} x^T \mathrm{Qx} + b^T \lambda - v^T (Q\,x + c + A^T \lambda) - z^T \lambda \\
\nabla_x W = Q\,x - Q\,v = 0 \\
\nabla_\lambda W = b - A\,v - z = 0 \qquad (a) \\
z^T \lambda = 0 \\
z \geqslant 0 \\
Q\,\overline{x} + c + A^T \overline{\lambda} = 0 \qquad (\Delta) \\
\lambda \geqslant 0
\end{aligned}
$$

Sono soddisfatte in $\overline{x}, \overline{\lambda}$ poiché per le ipotesi di KKT. Ora, devo trovare la soluzione x^* del primale. A tale scopo, sceiviamo KKT del primale e si deve avere:

$$
\begin{aligned}
\nabla_x L = 0 = -Q\,x + c + A^T \lambda \\
\lambda^T (A\,x - b) = 0 \\
\lambda \geqslant 0 \\
A\,x \leqslant b
\end{aligned}
$$

Da (a) è possibile ricavarsi $z = b - A\,v$ che, sostituito nella complementarietà del duale:

$$
\overline{\lambda}^T (b - A\,v) = 0 \longrightarrow z \geqslant 0 \,(\text{ammissibilità}) \longrightarrow A\,v \leqslant b (\text{ammissibilità primale di } v)
$$

Nota 93.

Quindi abbiamo trovato l'ammissibilà tra primale e duale con lo stesso λ.

Inoltre, so che:

$$
Q(\overline{x} - v) = 0 \longrightarrow Q\,\overline{x} = Q\,v
$$

Sostituendo quest'ultimo risultato in (Δ):

$$
Q\,v + c + A^T \overline{\lambda} = 0 \longrightarrow (x, v) \,\text{soddisfa KKT}
$$

A questo punto, abbiamo scritto KKT per il primale, in cui:

$$x^* = v \longrightarrow \text{La coppia } (x^*, \overline{\lambda}) \text{ soddisfa KKT del primale}$$

$$\longrightarrow x^* \text{ è soluzione ottima del primale} \rightarrow x^* = v \text{ è soluzione globale } Q\,\overline{x} - Q\,x^* = 0$$

$$\longrightarrow Q(\overline{x} - x^*) = 0 \qquad \square$$

Nota 94.

- $\overline{\lambda}$ viene trovato risolvendo il problema duale: ci si ricava la soluzione duale con cui risolvere il primale;
- Se $Q \succ 0, \overline{x} = x^*$ poiché $Q(\overline{x} - x^*) = 0 \longleftrightarrow \overline{x} - x^* = 0$

13 SVM Lineari

Il nostro obiettivo è quello di applicare la teoria della dualità della programmazione quadratica per risolvere il problema dell'iperpiano ottimo seguente:

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ y^i(w^T x^i + b) \geqslant 1 \quad i = 1, \ldots, l \quad (\lambda_i) \end{cases}$$

In questo tipo di problema è possibile applicare la dualità di Wolfe poiché la funzione obiettivo è convessa ed in particolare:

$$Q = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \succcurlyeq 0 \quad \text{semidefinita positiva}$$

Inoltre, i vincoli sono lineari.

Quindi iniziamo con lo scrivere la Lagrangiana di questo problema:

$$L(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \lambda_i (y^i (w^T x^i + b) - 1)$$

Il corrispondente duale di Wolfe, risulterà:

$$\max_{w,b,\lambda} \frac{1}{2} w^T w - \sum_{i=1}^{l} \lambda_i\, y^i\, w^T x^i - \sum_{i=1}^{l} \lambda_i y^i\, b + \sum_{i=1}^{l} \lambda_i$$

$$\nabla_w L(w, b, \lambda) = w - \sum_{i=1}^{l} \lambda_i\, y^i\, x^i = 0$$

$$\frac{\partial L(w, b, \lambda)}{\partial b} = -\sum_{i=1}^{l} \lambda_i\, y^i = 0$$

$$\lambda \geqslant 0$$

Dall'annullamento della Lagrangiana rispetto a b otteniamo che $\sum_{i=1}^{l} \lambda_i y^i\, b = 0$; mentre dall'annullamento della Lagrangiana rispetto a w otteniamo:

$$w = \sum_{i=1}^{l} \lambda_i\, y^i\, x^i$$

Sostituendo lìultima espressione nella funzione obiettivo, il problema può essere riscritto come:

$$\max \frac{1}{2}\sum_{i=1}^{l} \lambda_i\, y^i\, (x^i)^T \sum_{j=1}^{l} \lambda_i\, y^j\, x^j - \sum_{i=1}^{l} \lambda_i\, y^i\, (x^i)^T \sum_{j=1}^{l} \lambda_j\, y^j\, x^j + \sum_{i=1}^{l} \lambda_i =$$

$$= \max -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \lambda_i \lambda_j\, y^i\, y^j\, (x^i)^T x^j - \sum_{i=1}^{l} \lambda_i$$

Riscrivendo la funzione obiettivo in forma otteniamo il duale di Wolfe:

$$\begin{cases} -\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \lambda_i \lambda^j\, y^i\, y^j\, (x^i)^T x^j - \sum_{i=1}^{l} \lambda_i \\ \sum_{i=1}^{l} \lambda_i y^i = 0 \\ \lambda \geqslant 0 \quad i = 1, \ldots, l \end{cases}$$

A questo punto defiscono la forma matriciale di questo problema introducendo la matrice:

$$X = [y^1\, x^1 \ldots y^l\, x^l] \quad X \in \mathbb{R}^{n \times l}$$

In cui l'elemento $ij-$esimo risulterà nella forma:

$$[X^T X]_{ij} = y^i\, y^j\, (x^i)^T x^j$$

Quindi otteniamo il **duale in forma compatta**:

$$\min \frac{1}{2}\lambda^T X^T X\, \lambda - e^T \lambda$$

$$y^T \lambda = 0$$

$$\lambda \geqslant 0$$

in cui $e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$.

Nota 95.

$Q = X^T X \longrightarrow z^T Q\, z = z^T X^T X\, z = (X\, z)^T X\, z = \|X\, z\|^2 \geqslant 0 \longrightarrow Q \succcurlyeq 0$

Il problema è convesso in λ.

Quello a cui vogliamo rispondere è che:

- Il problema duale ammette soluzione?

  Per il teorema 86 so che il problema primale ammette soluzione unica, allora il problema duale ammette soluzione.

- Come si trovano w^* e b^* una volta che ho risolto il problema duale?

  Sia λ^* la soluzione del problema duale:

$$w^* = \sum_{i=1}^{l} \lambda_i^*\, y^i\, x^i \qquad \lambda^* \geqslant 0$$

Gli unici punti del Training Set che contribuiscono ad individuare w^* sono gli x^i per cui $\lambda^{i*} > 0$. I punti che soddisfano tale condizioni vengono detti **vettori di supporto.**

Risolvere il problema duale mi fornisce l'informazione sulla ridondanza del TS. Nel caso in cui $\lambda^{i*} > 0$, allora il vincolo $i-$esimo del primale è attivo all'uguaglianza e quindi il punto si trova sul margine.

- Come si ricava b^*?

  Per ricavare b^* si utilizza il teorema della dualità di Wolfe nel caso quadratico. In particolare λ^* è il moltiplicatore associato a w^*, b^* e quindi vale la complementarietà:

$$\lambda_i^*(y^i(w^{*T}x^i + b^*) - 1) = 0 \quad i = 1, \dots, l$$

  Se $\lambda_i^* > 0 \longrightarrow y^i(w^{*T}x^i + b^*) = 1$.Quindi:

$$b^* = \frac{1}{y^i} - w^{*T}x^i = \frac{1}{y^i} - \sum_{j=1}^{l} \lambda_j^* \, y^j \, (x^j)^T x^i$$

Quindi la **SVM** diventa:

$$y(x) = \text{sgn}(w^{*^T} x + b^*) = \text{sgn}\left(\sum_{i=1}^{l} \lambda_i^* \, y^i \, (x^i)^T \, x + b^* \right)$$

Riassumento, per trovare l'iperpiano ottimo (cioè quello con migliore capacità di generalizzazione tra quelli che separano i due insiemi A e B) posso risolvere o il problema primale o il duale e sono entrambi problemi di programmazione quadratica convessa.

Tutta la teoria funzione se A e B sono linearmente separabili. In caso contrario, l'insieme ammissibile è un insieme **vuoto** e quindi occorre rilassare i vincoli tramite l'aggiunta di variabili ausialiarie dette **variabili di slack**:

$$y^i \, (w^T x^i + b) \geqslant 1 - \xi_i \quad i = 1, \dots, l$$

$$\xi_i \geqslant 0$$

Quindi un punto x^i + **mal classificato** se i segni di y^i e $w^T x^i + b$ sono discordi. Allora:

$$y^i \, (w^T x^i + b) < 0$$

$$y^i \, (w^T x^i + b) \geqslant 1 - \xi_i \longleftrightarrow \xi_i \geqslant 1 - y^i(w^T x^i + b) \longrightarrow \xi_i > 1 \text{ se } x^i \text{ è } \textbf{mal classificato}$$

In particolare, $\sum_{i=1}^{l} \xi_i$ è un **upperbound** sul numero di punti mal classificati.

A fronte di ciò modifichiamo la funzione obiettivo del problema primale nel seguente modo;

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\begin{aligned} y^i \, (w^T x^i + b) &\geqslant 1 - \xi_i \quad i = 1, \dots, l \qquad \lambda_i \\ \xi_i &\geqslant 0 \qquad i = 1, \dots, l \qquad \mu_i \end{aligned}$$

Nota 96.

Il termine $C \sum_{i=1}^{l} \xi_i$ rappresenta una **penalità** sull'errore sul TS e $C > 0$ viene scelto in cross validation.

Scriviamo a questo punto la nuova Lagrangiana

$$L(w,b,\xi,\lambda,\mu)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\lambda_i\left(y^i\left(w^Tx^i+b\right)-1+\xi_i\right)-\sum_{i=1}^{l}\mu_i\xi_i$$

e il suo duale:

$$\max\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\lambda_i\left(y^i\left(w^Tx^i+b\right)-1+\xi_i\right)-\sum_{i=1}^{l}\mu_i\xi_i$$

$$\begin{aligned}
\Delta_w L(w,b,\xi,\lambda,\mu) &= w-\sum_{i=1}^{l}\lambda_i y^i x^i=0\\
\frac{\partial L}{\partial b}(w,b,\xi,\lambda,\mu) &= -\sum_{i=1}^{l}\lambda_i y^i=0\\
\frac{\partial L}{\partial \xi_i}(w,b,\xi,\lambda,\mu) &= C-\lambda_i-\mu_i \qquad i=1,\ldots,l\\
\lambda_i &\geqslant 0 \qquad i=1,\ldots,l\\
\mu_i &\geqslant 0 \qquad i=1,\ldots,l
\end{aligned}$$

Dall'annullamento della Lagrangiana rispetto a w e ξ possiamo ricavare:

$$w=\sum_{i=1}^{l}\lambda_i y^i x^i \qquad \mu_i=C-\lambda_i \qquad i=1,\ldots,l$$

Sostituendo queste due ultime espressioni, il problema diventa:

$$\max\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y^i y^j (x^i)^T x^j+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y^i y^j (x^i)^T x^j-\sum_{i=1}^{l}\lambda_i y^i b+\sum_{i=1}^{l}\lambda_i-\sum_{i=1}^{l}\lambda^i\xi^i-\sum_{i=1}^{l}(C-\lambda_i)\xi_i$$

Nota 97. I termini in rosso si semplificano o diventano nulli.

$$\max-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y^i y^j (x^i)^T x^j+\sum_{i=1}^{l}\lambda_i$$

Riscrivendolo in forma di minimo:

$$-\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y^i y^j (x^i)^T x^j-\sum_{i=1}^{l}\lambda_i$$

$$\begin{aligned}
\sum_{i=1}^{l}\lambda^i y^i &= 0\\
\lambda_i &\geqslant 0\\
C-\lambda_i &\geqslant 0 \quad i=1,\ldots,l
\end{aligned}$$

Gli ultimi due vincoli vengono riscritti come $0\leqslant\lambda_i\leqslant C \quad i=1,\ldots,l$

Alla fine il duale diventa:

$$-\min \quad \frac{1}{2}\lambda^T X^T X \lambda - e^T \lambda$$

$$y^T \lambda = 0$$

$$0 \leqslant \lambda_i \leqslant C \quad i = 1, \dots, l$$

A questo punto supponiamo di risolvere il duale λ^*, vogliamo capire come si trovano w^*, b^* del primale. A tale scopo:

$$w^* = \sum_{i=1}^{l} \lambda_i^* \, y^i \, x^i$$

Importante 98. I vettori di supporto sono gli x^i per cui $\lambda_i^* > 0$.

In aggiunta, b^* si trova tramite la complementarieta. In particolare, (λ^*, μ^*) dove $\mu_i^* = C - \lambda_i^*$ sono i moltiplicatori associati alla soluzione primale. Allora questi moltiplicatori devono soddisfare la complementarietà:

$$\lambda_i^* (y^i \, ({w^*}^T x^i + b^*) - 1 + \xi_i^*) = 0 \quad i = 1, \dots, l$$

$$\mu_i^* \, \xi_i^* = (C - \lambda_i^*) \, \xi_i^* = 0 \quad i = 1, \dots, l$$

$$\xi_i^*(C - \lambda_i^*) = 0 \longrightarrow \text{Se } \lambda_i^* < C \longrightarrow \xi_i^* = 0 \longrightarrow x^i \text{ è } \textbf{ben calssificato}$$

Per trovare b^* scelgo $i: 0 < \lambda_i^* < C$. Allora $\xi_i^* = 0 \longrightarrow y^i({w^*}^T x^i + b^*) = 1$.

I punti x^i per cui $\lambda_i^* = C$ non posso calcolare ξ_i^*. Allora sono candidati ad essere malclassificati. Infine, la nostra macchina SVN sarà del tipo:

$$y(x) = \text{sgn}({w^*}^T x + b^*)$$

14 SVM Non Lineari

Per alcuni insiemi di dati una superficie di separazione lineare non cattura correttamente la classificazion. A tale scopio, passiamo dallo spazio di input a uno spazio di dimensione maggiore, detto **feature space.**

Quello che non era linearmente separabile nello spazio di input lo può diventare nello spazio delle feature. La superficie che si ottiene è lineare nello spazio delle feature, ma non lineare nello spazio di partenza. In partica, dobbiamo utilizzare il **kernel kick**.

14.1 Kernel trick

La funzione kernel calcola ul prodotto scalare di due trasformazioni dell'input. L'idea è quella di pasare da:

$$x \, (\text{spazio di input}) \longrightarrow \Phi(x) = \{\Phi(x_1), \Phi(x_2), \dots\} \, (\text{spazio delle feature})$$

L'iperpiano di sperazione viene cercato nello spazio delel feature e la corrispondente macchina SVM diventa:

$$y(x) = \text{sgn}(w^T \Phi(x) + b)$$

Lo spazio di arrivo può avere dimensione infinita.

La funzione $\Phi(x)$ non viene calcolata esplicitramente, ma si sfrutta il fatto che x, nella SVM lineari non compare mai da sola, ma sempre come prodotto $x_i^T x_j$ oppure $x^T x^i$. Quindi, mi serve calcolare $\Phi(x_i)^T \Phi(x_j)$ o $\Phi(x)^T \Phi(x^i)$.

Di conseguenza, il problema di ottimizzazione nel duale diventa:

$$\min \frac{1}{2} \sum_i \sum_j y^i y^j \lambda_i \lambda_j \Phi(x^i)^T \Phi(x^j) - \sum_{i=1}^{l} \lambda_i$$

$$\begin{aligned} \sum_i \lambda_i y^i &= 0 \\ 0 \leqslant \lambda_i &\leqslant C \end{aligned}$$

La b^* si calcola scegliendo $0 \leqslant \lambda_i^* \leqslant C \longrightarrow b^* = \frac{1}{y^i} - \sum_{j=1}^{l} y^j \lambda_j^* \Phi(x^j)^T \Phi(x^i)$

La SVM diventa quindi:

$$y(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \lambda_i^* y^i \, \Phi(x^j)^T \Phi(x^i) + b^* \right)$$

14.2 Funzioni Kernel

Dato un insieme $X \subseteq \mathbb{R}^n, K: X \times X \to \mathbb{R}$ è una **funzione Kernel** se:

$$k(x, y) = \Phi(x)^T \Phi(y)$$

In cui $\Phi: \mathbb{R}^n \to H$ spazio Euclideo.

Nota 99.

- Se K è un Kernel allora $\forall x^1, \ldots, x^l \in X$ definisco:

  $$\mathrm{K} = [K(x^i, x^j)] \quad i = 1, \ldots, l; j = 1, \ldots, l$$

  è simmetrica e semidefinita positiva, quindi non si perdono le proprietà di convessità.

  $$K = \begin{pmatrix} k(x^i, x^j) & \ldots & k(x^1, x^l) \\ \vdots & \ddots & \vdots \\ k(x^l, x^1) & \ldots & k(x^l, x^l) \end{pmatrix} \succcurlyeq 0$$

- Si usa la funzione Kernel nel duale:

  $$\min \frac{1}{2} \sum_i \sum_j y^i y^j \lambda_i \lambda_j K(x^i, x^j) - \sum_{i=1}^{l} \lambda_i$$

  $$\begin{aligned} \sum_i \lambda_i y^i &= 0 \\ 0 \leqslant \lambda_i &\leqslant C \end{aligned}$$

- Il problema rimane convesso, si risolve il duale trovando λ^* e si trova b^* per complementarietà:

$$b^* = \frac{1}{y^i} - \sum_{j=1}^{l} y^j \lambda_j^* k(x^i, x^j)$$

La SVM diventa:

$$y(x) = \text{sgn}\left(\sum_{i=1}^{l} \lambda_i y^i k(x^i, x) + b^* \right)$$

I Kernel più utilizzati sono:

1. **Lineare:** $K(x, y) = x^T y$
2. **Polinomiale:** $k(x, y) = (x^T y + \gamma)^p \quad \gamma \geqslant 0, p \geqslant 1$ (Scelti in cross validation)
3. **Gaussiano:** $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \quad \sigma > 0$
4. **Tangente Iperbolica:** $k(x, y) = \tanh(\beta x^T y + \gamma)$

14.3 Kernel Polinomiale

$$k(x, y) = (x^T y + \gamma)^p = \Phi(x)\,\Phi(y) \quad \Phi = \mathbb{R}^n \longrightarrow H$$

La dimensione minima di H è $\binom{n+p-1}{n} = \frac{(n+p-1)!}{n!\,(p-1)!}$

14.4 Kernel Gaussiano

$$k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \qquad \gamma = \frac{1}{2\,\sigma^2} \longrightarrow k(x, y) = e^{-\gamma \|x-y\|^2}$$

Tramite un Kernel Gaussiano, nello spazio delle feature, tutto è linearmente separabile poiché lo spazio delle feature ha dimensione infinita. (Dato che con un semplice esempio possiamo esprimere l'esponenziale tramite una serie infinita)

Data questa proprietà del Kernel Gaussiano possiamo quindi affermare che la teoria di Vapnik rimane valida e quindi assicurare l'esistenza dell'iperpiano ottimo. Tuttavia, questa scelta non risulta sempre ottimale poiché il costo computazionale potrebbe risultare oneroso e si cade facilmente nel caso di overfitting dei dati.

15 Working Set

Dato un Training Set:

$$\text{TS} = \{(x^i, y^i), x^i \in \mathbb{R}^n, y^i \in \{-1, 1\}\, i = 1, \ldots, k\}$$

Trovare l'iperpiano a massimo margine vuol dire risolvere:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$y^i (w^T x^i b) \geqslant 1 - \xi_i \quad i = 1, \ldots, k\,(\lambda_i)$$

$$\xi_i \geqslant 0$$

Equivalentemente possiamo risolvere il duale:

$$\min \frac{1}{2} \sum_i \sum_j y^i y^j (x^i)^T x^j - \sum_{i=1}^{l} \lambda_i$$

$$\sum_{i=1}^{l} \lambda_i y^i = 0$$

$$0 \leqslant \lambda_i \leqslant C$$

Che si risolve ottenendo:

$$w^* = \sum_{i=1}^{l} \lambda_i^* y^i x^i$$

e per complementarietà:

$$b^*$$

La SVM risultante sarà:

$$y(x) = \text{sgn}(w^{*T} x + b^*)$$

Così facendo si produce una SVM lineare.

Nel caso delle SVM non lineare si usa il kernel trick e si risolve il problema:

$$\min \frac{1}{2} \sum_i \sum_j y^i y^j K(x^i, y^i) - \sum_{i=1}^{l} \lambda_i$$

$$\sum_{i=1}^{l} \lambda_i y^i = 0$$

$$0 \leqslant \lambda_i \leqslant C$$

La SVM risultante sarà del tipo:

$$y(x) = \text{sgn}\left(\sum_{i=1}^{l} \lambda_i^* y^i \, K(x^i, x) b^* \right)$$

Indipendentemente dal Kernel che si utilizza, il problema duale è nella forma:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \, \alpha - e^T \alpha$$

$$y^T \alpha = 0 \quad Q \succcurlyeq 0$$

$$0 \leqslant \alpha \leqslant C$$

$Q_{ij} = y^i \, y^j K(x^i, x^j)$ con $Q = Q^T \quad Q \in \mathbb{R}^{l \times l}$

Talvolta, Q può essere talmente gfrande da avere problemi di memoria allora si trova in cache un sottoinsieme di colonne. Allora si utilizzano delle **tecniche di decomposizione**: si ottimizza ogni volta rispetto ad un sottoinsieme di variabili detto **working set** dipenendete dal tipo di kernel utilizzato. In particolare:

Si partiziona $\alpha = \begin{pmatrix} \alpha\, w \\ \alpha\, \overline{w} \end{pmatrix} \quad \overline{w} = \{1, \ldots, n\} / W$ con le seguenti proprietà:

- $W \cup \overline{W} = \{1, \ldots, l\}$
- $W \cap \overline{W} = \varnothing$

Parto da α_k ammissibile: $y^T \alpha_k = 0$ con $0 \leqslant \alpha_k \leqslant C \quad i = 1, \ldots, l$

Calcolo α_W^{k+1} fissando tutte le variabili tranne quelle in W:

$$\alpha_{\overline{W}} = \alpha_{\overline{W}}^k$$

$$Q = \begin{pmatrix} Q_{WW} & Q_{W\overline{W}} \\ Q_{\overline{W}W} & Q_{\overline{W}\overline{W}} \end{pmatrix}$$

Il problema da risolvere diventa:

$$\min_{\alpha w} f(\alpha\, w, \alpha_{\overline{W}}^k) = \frac{1}{2} \begin{pmatrix} \alpha_W^T & \alpha_{\overline{W}}^{k^T} \end{pmatrix} \begin{pmatrix} Q_{WW} & Q_{W\overline{W}} \\ Q_{\overline{W}W} & Q_{\overline{W}\overline{W}} \end{pmatrix} \begin{pmatrix} \alpha_W \\ \alpha_{\overline{W}}^k \end{pmatrix} - e^T \alpha_W - e^T \alpha_{\overline{W}} =$$

$$= \frac{1}{2} \alpha_W^T Q_{WW} \alpha_W + \frac{1}{2} \alpha_{\overline{W}}^k Q_{\overline{W}\overline{W}} \alpha_{\overline{W}}^k + \alpha_{\overline{W}}^k Q_{\overline{W}W} \alpha_W - e^T \alpha_W =$$

$$= \frac{1}{2} \alpha_W^T Q_{WW} \alpha_W + c^T \alpha_w \qquad c = Q_{W\overline{W}}\, \alpha_{\overline{W}}^k - e$$

$\alpha_W^{k+1} = \arg\min_{\alpha_W} \frac{1}{2} \alpha_W^T Q_{WW} \alpha_W + c^T \alpha_w$.

$$y_W^T \alpha_W + y_{\overline{W}}^T \alpha_{\overline{W}}^k = 0 \rightarrow y_W^T \alpha_W = -y_{\overline{W}}^T \alpha_{\overline{W}}^k$$

$$0 \leqslant \alpha_W \leqslant C$$

La struttura del problema rimane inviariata ma la dimensione diventa $|W|$ più piccolo.

L'algoritmo implementativo è il seguente.

Algoritmo 1

DATI
$\alpha^o (\alpha^o = 0)$
INIT $k = 0$
WHILE criterio di arresto non è soddisfatto

1. Seleziono tramite WSS un insieme W^k
2. Poni $W = W^k$ e risolvi:
$$\alpha_W^* = \arg\min \frac{1}{2} \alpha_W^T Q_{WW} \alpha_W + c^T \alpha_w$$
$$y_W^T \alpha_W = -y_{\overline{W}}^T \alpha_{\overline{W}}^k$$
$$0 \leqslant \alpha_W \leqslant C$$
3. Poni:
$$\alpha_i^{k+1} = \begin{cases} \alpha_i^* & \text{se } i \in W \\ \alpha_i^k & \text{se } i \in \overline{W} \end{cases}$$

4. $k = k + 1$
   END

Importante 100. La convergenza dipende dalla selezione del WS.

Devo scegliere la cardinalità di W, alora se è piccolo il sottoproblema facile, ma si necessitano di più iterazioni e quindi la convergenza risulta più lento. La cardinalità minima è 2

Se $|W| = 2$ allora si hanno algoritmi di tipo SMO (Sequential Minimization Ottimization) e la soluzione è analitica, ma non si ha necessità di algoritmi e l'iterazione è molto veloce.

In generale non si supera $|W| \leqslant 10$, se no l'iterazione diventa molto costota. Lo svantaggio degli algoritmi a di decomposizione è la lentezza dato il grande numero di iterazioni e non sono molto precisi.

Teorema 101. *Riformulazione KKT*

$$\alpha^* \text{ ammissibile ottimo globale} \iff \exists \lambda^* : \nabla_i f(\alpha^*) + \lambda^* y^i = \begin{cases} \geqslant 0 \text{ se } i \in L(\alpha^*) \\ = 0 \text{ se } i \notin L(\alpha^*) \cup U(\alpha^*) \\ \leqslant 0 \text{ se } i \in U(\alpha^*) \end{cases}$$

In cui:

$$L(\alpha^*) = \{i : \alpha_i^* = 0\}$$

$$U(\alpha^*) = \{i : \alpha_i^* = C\}$$

Dimostrazione.

$\Longrightarrow$Se vale KKT allora vale il teorema

Supponiamo che $i \in L(\alpha^*)$ allora $\alpha_I^* = 0$. La sostiuiamo nella complementarietà $\hat{\xi}_i = 0$. A questo punto sostituisco nell'annullamento del gradiente del Lagrangiano e si ottiene:

$$\nabla_i f(\alpha^*) + \lambda^* y^i - \xi_i = 0$$

$$\nabla_i f(\alpha^*) + \lambda^* y^i = \xi_i = 0$$

Supponiamo che $i \in U(\alpha^*) \longrightarrow \alpha_i^* = C$. La sostituiamo nella complementarietò $\xi_i = 0$. A questo punto sostituisco nell'annullamento del gradiente del Lagrangiano:

$$\nabla_i f(\alpha^*) + \lambda^* y^i + \hat{\xi}_i = 0$$

$$\nabla_i f(\alpha^*) + \lambda^* y^i = -\hat{\xi}_i \leqslant 0$$

Supponiamo infine che $i \notin L(\alpha^*) \cup U(\alpha^*) \longrightarrow 0 < \alpha_i^* < C$. Dalla complementarietà $\xi_i = 0$ $\hat{\xi}_i = 0$. A questo punto sostituiamo nell'annullamento del gradiente della Lagrangiana:

$$\nabla_i f(\alpha^*) + \lambda^* y^i = 0$$

$\Longleftarrow$Se vale il teorema allora vale KKT

Dobbiamo generare dei moltiplicatori che devono soddisfare KKT e ξ_i e $\hat{\xi}_i$. Nel caso in cui:

- $i \in L(\alpha^*) \longrightarrow \alpha_i^* = 0$ per soddisfare la complementarietà prendo:

$$\hat{\xi}_i = 0 \longrightarrow \hat{\xi}_i(\alpha_i - C) = 0$$

Devo scegliere $\xi_i \geqslant 0$ per soddisfare l'annullamento del gradiente della lagrangiana ed in particolare scelgo:

$$\xi_i = \nabla_i f(\alpha^*) + \lambda^* y^i \geqslant 0 \,\text{Per ipotesi}$$

- $i \in U(\alpha^*) \rightarrow \alpha_i = C$ per soddisfare la complementarietà $\xi_i = 0$

$$(\alpha_i - C)\hat{\xi}_i = 0$$

è soddisfatta perché $\alpha_i^* = C$. Devo scegliere $\hat{\xi}_i \geqslant 0$ tale che sia sodisfatto l'annullamento del gradiente della lagrangiana e quindi si ottiene:

$$\hat{\xi}_i = -\nabla_i f(\alpha^*) + \lambda^* \, y^i \geqslant 0$$

- $i \notin L(\alpha^*) \cup U(\alpha^*) \rightarrow 0 < \alpha_i^* < C$. Quindi:

$$\xi_i = \hat{\xi}_i = 0 \rightarrow \xi_i \, \alpha_i = 0 \quad \hat{\xi}_i(\alpha_i - C) = 0 e \,\text{per ipotesi}\, \nabla_i L = 0$$

□

Nota 102.

Dal teorema si possono definire i seguenti insiemi:

$$L(\alpha^*) = L^+(\alpha^*) \cup L^-(\alpha^*)$$

In cui:

$$L^+(\alpha^*) = \{i : \alpha_i^* = 0 e y^i = 1\}$$

$$L^-(\alpha^*)\{i : \alpha_i^* = 0 e y^i = -1\}$$

Inoltre:

$$U(\alpha^*) = U^+(\alpha^*) \cup U^-(\alpha^*)$$

In cui:

$$U^+(\alpha^*) = \{i : \alpha_i^* = C e y^i = 1\}$$

$$U^-(\alpha^*)\{i : \alpha_i^* = \text{Ce } y^i = -1\}$$

Pontendo così riscrivere le condizioni del teorema come segue:

$$\lambda^* y^i = \begin{cases} \geqslant -\nabla_i f(\alpha^*) & i \in L(\alpha^*) \\ = -\nabla_i f(\alpha^*) & i \notin L(\alpha^*) \cup U(\alpha^*) \\ \leqslant -\nabla_i f(\alpha^*) & i \in U(\alpha) \end{cases}$$

$$\lambda^* = \begin{cases} \geqslant -\frac{\nabla_i f(\alpha^*)}{y^i} & i \in L^+(\alpha^*) \cup U^-(\alpha^*) \\ = -\frac{\nabla_i f(\alpha^*)}{y^i} & i \in L(\alpha^*) \cup U(\alpha^*) \\ \leqslant -\frac{\nabla_i f(\alpha^*)}{y^i} & i \in L^-(\alpha^*) \cup U^+(\alpha^*) \end{cases}$$

Oltre a questi parizionamenti possiamo definire i seguenti insiemi aggiuntivi:

$$R(\alpha) = L^+(\alpha^*) \cup U^-(\alpha^*) \cup \{i: 0 < \alpha_i < C\}$$

$$S(\alpha) = L^-(\alpha^*) \cup U^+(\alpha^*) \cup \{i: 0 < \alpha_i < C\}$$

di cui:

$$R(\alpha) \cup S(\alpha) = \{1, \ldots, l\}$$

$$R(\alpha) \cap S(\alpha) = \{i: 0 < \alpha_i < C\}$$

Teorema 103.

$$\left.\begin{array}{l} \alpha^* \text{ ammissibile} \\ \\ R(\alpha^*) \, o \, S(\alpha^*) \text{ non vuoto} \end{array}\right. \implies \alpha^* \text{ è ottimo}$$

Dimostrazione.

Supponiamo $R(\alpha^*) = \varnothing \rightarrow S(\alpha^*) = \{1, \ldots, l\}$ e $R(\alpha^*) \cap S(\alpha^*) = \varnothing$,

Allora $\nexists i$ tale che; $0 < \alpha_i < C \longrightarrow S(\alpha^*) = L^-(\alpha^*) \cup U(\alpha^*) = \{1, \ldots, l\}$

La condizione di ottimo diventa:

$$\lambda^* \leqslant -\frac{\nabla_i f(\alpha^*)}{y^i} \quad i = 1, \ldots, l$$

$$\rightarrow \lambda^* = \min_{i=1,\ldots,l} \left\{ -\frac{\nabla_i f(\alpha^*)}{y^i} \right\}$$

Soddisfa le condizioni di ottimo globale.

Al contrario se $S(\alpha^*) = \varnothing \quad R(\alpha^*) = \{1, \ldots, l\}$ e $R(\alpha^*) \cap S(\alpha^*) = \varnothing$

Allora $\nexists i: 0 < \alpha^* < C$:

$$R(\alpha^*) = L^+(\alpha^*) \cup U^-(\alpha^*)$$

La condizione di ottimo diventa:

$$\lambda^* \geqslant -\frac{\nabla_i f(\alpha^*)}{y^i} \quad i = 1, \ldots, l$$

$$\lambda^* = \max_{i=1,\ldots,l} \left\{ -\frac{\nabla_i f(\alpha^*)}{y^i} \right\}$$

Il che soddisfa le condizioni di ottimo. □

Teorema 104. *Condizione di ottimo*

α^* *ammissibile è ottima* $\iff \max_{i=1,\ldots,l} \left\{ -\frac{\nabla_i f(\alpha^*)}{y^i} \right\} \leqslant \min_{i=1,\ldots,l} \left\{ -\frac{\nabla_i f(\alpha^*)}{y^i} \right\}$

16 SMO

SMO sta per sequential minimization optimization ed in questi tipi di algoritmi soono interessato ad avere un Working Set di cardinalità 2 di cui muovo due variabili. In particolare, anche in questa tipologia di algoritmi effettuiamo decomposizione e vogliamo cercare di caratterizzare le direzioni ammissibili su cui devo decomporre. A tale scopo, definiamo l'insieme:

$$F = \{\alpha \in \mathbb{R}^l : y^T\alpha = 0, 0 \leqslant \alpha_i \leqslant C, i = 1, \dots, l\}$$

Ho vincoli lineari, uno di uguaglianza e 2*l vincoli di box:

$$a_i^T\alpha_i = b_i \longrightarrow a_i^T(a+d) = b_i \longleftrightarrow a_i^T\alpha + a_i^T d = b_i \longrightarrow a_i^T d = 0$$

In questo caso:

$$y^T\alpha = 0 \longrightarrow d \text{ è ammissibile} \longleftrightarrow y^T d = 0$$

Se mi trovo in α ammissibile e $\alpha_i = 0 \longrightarrow d_i \geqslant 0$

Se mi trovo in α ammissibile e $\alpha_i = C \longrightarrow d_i \leqslant 0$

Ricapitolando in α ammissibile l'insieme delle direzioni ammissibile è definito come:

$$\{d \in \mathbb{R}^l : y^T d = 0, d_i \geqslant 0 \quad \forall i \in L(\overline{\alpha}), d_i \leqslant 0, \forall U(\overline{\alpha})\}$$

Teorema 105.

Sia $\overline{\alpha}$ ammissibile e sia $\{(i,j) \in \{1, \dots, l\}\} \quad i \neq j$. Definisco $d^{i,j}$ come:

$$d_h^{i,j} = \begin{cases} \frac{1}{y^i} & \text{se } h = i \\ 0 & \text{altirmenti} \\ -\frac{1}{y^j} & \text{se } h = j \end{cases}$$

Allora:

- *d ammissibile $\Longleftrightarrow i \in R(\overline{\alpha}) \quad e\, j \in S(\overline{\alpha})$*
- *$d^{i,j}$ è di discesa per f in $\overline{\alpha}$ se:*

$$\frac{\nabla_i f(\alpha^*)}{y^i} - \frac{\nabla_j f(\alpha^*)}{y^j} < 0$$

Dimostrazione.

- Se d ammissibile $\longrightarrow y^T d = 0$ banale.

  Se $d^{i,j}$ ammissibile $\longrightarrow i \in R(\overline{\alpha}) \quad e \quad j \in S(\overline{\alpha})$. Suppongo per assurdo che non si vero:

  $$j \in R(\overline{\alpha}) = L^+(\overline{\alpha}) \cup U^-(\overline{\alpha}) \cup \{i : 0 < \alpha_i < C\}$$

  Poiché $j \in S(\overline{\alpha}) \longrightarrow j \notin \{i : 0 < \alpha_i < C\} \longrightarrow j \in L^+(\overline{\alpha})$.

  Quindi se $j \in L^+(\overline{\alpha}) \rightarrow d^{i,j} = -\frac{1}{y^i} \leqslant 0$

  Il che è un assurco perché per essere ammissibile al lower bound deve essere $\geqslant 0$.

- $d^{i,j}$ è di discesa$\longleftrightarrow \nabla f(\overline{\alpha})^T d^{i,j} \leqslant 0$

  Quindi:

$$\nabla f(\overline{\alpha})^T d^{i,j} = \frac{\nabla_i f(\alpha^*)}{y^i} - \frac{\nabla_j f(\alpha^*)}{y^j} < 0$$

□

Definisco:

- $I(\alpha) = \left\{ i \colon \arg\max_{i \in R(\alpha)} \left\{ -\frac{\nabla_i f(\alpha)}{y_i} \right\} \right\}$
- $J(\alpha) = \left\{ j \colon \arg\max_{j \in S(\alpha)} \left\{ -\frac{\nabla_i f(\alpha)}{y_i} \right\} \right\}$

Se α non è ottimo$\rightarrow I(\alpha)$ e $J(\alpha)$ sono ben definiti e non vuoti.

Se α non è ottimo, la condizione di ottimalità risulta violata:

$$\max_{i \in R(\alpha)} \left\{ -\frac{\nabla_i f(\alpha^*)}{y^i} \right\} \leqslant \min_{j \in J(\alpha)} \left\{ -\frac{\nabla_j f(\alpha^*)}{y^j} \right\}$$

Suppondiamo di scegliere $i \in I(\alpha)$ e $j \in J(\alpha)$ e considere una direzione d^{ij}. Allora:

1. d^{ij} è ammissibile poiché $i \in I(\alpha) \subseteq R(\alpha)$ e $j \in J(\alpha) \subseteq S(\alpha)$
2. d^{ij} è di discesa ed è quella con derivata direzionale più bassa.

$i \in I(\alpha)$

$$-\frac{\nabla_i f(\alpha^*)}{y^i} \geqslant -\frac{\nabla_t f(\alpha^*)}{y^t} \qquad \forall t \in R(\alpha) \quad (1)$$

$j \in J(\alpha)$

$$-\frac{\nabla_j f(\alpha^*)}{y^j} \leqslant -\frac{\nabla_s f(\alpha^*)}{y^s} \quad \forall s \in S(\alpha) \quad (2)$$

Moltplicando (1) per -1:

$$\frac{\nabla_i f(\alpha^*)}{y^i} \leqslant \frac{\nabla_t f(\alpha^*)}{y^t} \quad (3)$$

Sommando (2) e (3):

$$\frac{\nabla_i f(\alpha^*)}{y^i} - \frac{\nabla_j f(\alpha^*)}{y^j} \leqslant \frac{\nabla_t f(\alpha^*)}{y^t} - \frac{\nabla_s f(\alpha^*)}{y^s} \quad \forall t \in R(\alpha) \quad \forall s \in S(\alpha)$$

Dalla condizione di ottimo violata, si ottiene:

$$-\frac{\nabla_i f(\alpha^*)}{y^i} > -\frac{\nabla_j f(\alpha^*)}{y^j} \longrightarrow \frac{\nabla_i f(\alpha^*)}{y^i} - \frac{\nabla_j f(\alpha^*)}{y^j} < 0$$

è la massima direzione di discesa che viola maggiormente le condizioni di ottimo.

Algoritmo

Dati $\alpha^\circ = 0, \nabla f(\alpha^\circ) = -e$

Init $k=0$
While criterio di arresto non soddisfatto

1. Determina Working Set WS:$i \in I(\alpha^k), j \in J(\alpha^k)$ e poni $W_k = \{i, j\}$
2. Determina analiticamente una soluzione α_i^*, α_j^* del problema ristretto:

$$\min_{\alpha_i,\alpha_j} q(\alpha_i, \alpha_j) = \frac{1}{2}\begin{pmatrix} \alpha_i & \alpha_j \end{pmatrix}\begin{pmatrix} q_{ii} & q_{ij} \\ q_{ji} & q_{jj} \end{pmatrix}\begin{pmatrix} \alpha_i \\ \alpha_j \end{pmatrix} + \sum_{h\neq i,j} (q_{ih}\,\alpha_i + q_{jh}\,\alpha_j)\,\alpha_h^k - \alpha_i - \alpha_j + \text{cost}$$

$$y^i\,\alpha_i + y^j\,\alpha^j = -\sum_{h\neq i,j} y^k\,\alpha_h^k$$

$$0 \leqslant \alpha_i \leqslant C \quad 0 \leqslant \alpha_j \leqslant C$$

3. Poni:

$$\alpha_h^{k+1} = \begin{cases} \alpha_i^* & h = i \\ \alpha_j^* & h = j \\ \alpha^* & h \neq \{i, j\} \end{cases}$$

4. Poni $\nabla f(\alpha^{k+1}) = Q\,\alpha^{k+1} - e = Q\,\alpha^{k+1} \pm Q\,\alpha^k - e = Q\,\alpha^k - e + Q(\alpha^{k+1} - \alpha^k) =$

$$= \nabla f\,\alpha^k + Q_i(\alpha^{k+1} - \alpha_i^k) + Q_j\,(\alpha^{k+1} - \alpha_j^k)$$

5. $k = k + 1$

END

Teorema 106.

Se $Q \succcurlyeq 0 \quad Q = Q^T$ *ogni punto limite delle sequenza* $\{\alpha^k\}$ *è soluzione del duale.*

16.1 Criterio di Arresto

La condizione di ottimo è:

$$\max_{i\in R(\alpha)} \left\{-\frac{\nabla_i f(\alpha^*)}{y^i}\right\} \leqslant \min_{i\in S(\alpha)} \left\{-\frac{\nabla_i f(\alpha^*)}{y^i}\right\}$$

Per rilassare questa condizione imponiamo:

$$m(\alpha) = \begin{cases} \max_{i\in R(\alpha)} \left\{-\frac{\nabla_i f(\alpha^*)}{y^i}\right\} & \text{se } R(\alpha) \neq \varnothing \\ -\infty & \text{altrimenti} \end{cases}$$

$$M(\alpha) = \begin{cases} \min_{i\in R(\alpha)} \left\{-\frac{\nabla_i f(\alpha^*)}{y^i}\right\} & \text{se } S(\alpha) \neq \varnothing \\ \infty & \text{altrimenti} \end{cases}$$

Quindi, α è soluzione ottima $\Longleftrightarrow m(\alpha) \leqslant M(\alpha)$.

Mentre la sequenza $\{\alpha^k\}$ convergente all'ottimo si potrebbe avere $m(\alpha) \geqslant M(\alpha)$.

Un criterio di arresto utilizzato è:

$$m(\alpha^k) \leqslant M(\alpha^k) + \varepsilon \qquad \varepsilon > 0$$

Si potrebbe avere:

$$\lim_{k\to\infty} m(\alpha^k) \neq m(\alpha^*) \text{ di norma discontinua}$$

La convergennza viene garantita teoricamente aggiungendo ipotesi sulla struttura della matrice Q.

17 Riepilogo SVM

Nel problema delle SVM si vuole trovare l'iperpiano a massimo margine. Questa scelta è ottima se i due insiemi risultano linearmente separabili. In particolare, il problema primale è il seguente:

- Se si vuole minimizzare l'errore in base alla **L1-Loss (modulo dell'errore)**

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$y^i(w^T x^i b) \geqslant 1 - \xi_i \quad i = 1, \ldots, k\,(\lambda_i)$$

$$\xi_i \geqslant 0$$

- Se si vuole minimizzare l'error ein base alla **L2-Loss (modulo dell'errore al quadrato)**

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i^2$$

$$y^i(w^T x^i b) \geqslant 1 - \xi_i \quad i = 1, \ldots, k\,(\lambda_i)$$

$$\xi_i \geqslant 0$$

Il suo duale sia se si tratta di un problema di L1-LOSS sia che si tratta di un problema con L2-Loss, è il seguente:

$$\min \frac{1}{2}\lambda^T Q\,\lambda - e^T\lambda_i$$

$$\sum_{i=1}^{l} \lambda_i y^i = 0$$

$$0 \leqslant \lambda_i \leqslant C$$

In cui:

$$Q_{ij} = \begin{cases} y^i y^j (x^i)^T x^j & \underline{\text{Kernel Lineare} \rightarrow \text{Ipepriano ottimo nello spazio di input}} \\ y^i y^j K(x^i x^j) & \underline{\text{kernel Oppurtuno} \rightarrow \text{SVM non lineare} \rightarrow \text{Iperpiano spazio}} \text{ delle features} \end{cases}$$

Si possono utilizzare pesi differenti in base alla classe $\longrightarrow \frac{1}{\#\text{Istante TS della classe } i}$

La funzione obiettivo diventa:

$$C\sum_{i=1}^{n} \xi_i \longrightarrow C\sum_{i \in A} w_a \xi_i + C\sum_{j \in B} w_b \xi_j$$

Importante 107. Il parametro C viene scelto in cross valdiation, ma di solito viene impostato a $C \approx \frac{1}{10}$. Nel caso in cui diventasse troppo grande per evitare overfitting, il problema si complica.

18 ν-Classification

Nella $\nu -$ Classification si risolve il problema:

$$\begin{cases} \min \frac{1}{2}\|w\|^2 - \nu\,\rho + \frac{1}{2}\sum \xi_i \\ y^i(w^T x^i + b) \geqslant \rho - \xi_i \quad \xi_i = 1, \ldots, l \\ \xi_i \geqslant 0 \quad \rho \geqslant 0 \quad i = 1, \ldots, l \end{cases}$$

Se si aumenta C diminusice ν e viceversa.

Se $\xi_i = 0 \longrightarrow y^i\,(w^T x^i + b) \geqslant \rho$. La quantità ρ rappresenta il **margine** e vale:

$$\frac{2\,\rho}{\|w\|}$$

Nel caso in cui $\rho > 0$, all'ottimo:

1. $\nu \geqslant$ frazione di errore nel margine (punti che si trovano sul margine)
2. $\nu \leqslant$ frazione dei vettori di supporto del TS
3. Con probabilità 1 al limite il punto 1 e 2 sonouguali.

Se $\rho > 0$ all'ottimo è equivalente alla $C -$ Classification con $C = \frac{1}{2\,\rho}$.

Importante 108.

In particolare $\nu \in [\nu_{\min}, \nu_{\max}]$ o il duale diventa inammissibile:

$$\begin{cases} \min \frac{1}{2}\,\alpha^T Q\,\alpha \\ y^T \alpha > 0 \quad e^T \alpha \geqslant \nu \\ 0 \leqslant \alpha_i \leqslant \frac{1}{l} \end{cases}$$

$$\nu_{\max} = \frac{2\min\left(|A|, |B|\right)}{l}$$

Con:

- $A = \{i = 1, \ldots, l \quad y_i = 1\}$
- $B = \{i = 1, \ldots, l \quad y_i = -1\}$

Se si utilizzasse LIBSVM l'algoritmo di tipo SMO sceglie $i \in I(\alpha)$ e j guardando le condizioni del secondo ordine. Per velocizzare l'algoritmo si usano le tecniche di **Shrinking**;

1. Se $\lambda^i = 0\,(\alpha_i = 0)$ per tante iterazioni,allora la fisso a 0 riducendo conseguentemente il numero di variabili;
2. Se $\lambda^i = C$ per tante iterazioni allora la fisso a C;

Tuttavia si deve ocmunque verficare che le condizioni di ottimo siano soddisfatte. In aggiunta il costo generico di una iterazione:

$$O(|W|\,l)$$

Utilizzando un Kernel Lineare (es. LIBLINEAR) il costo delll'iterazione si abbatte.

Ricordando un problema con Kernel Lineare e valudanto il suo duale, se N è molto grande (molte features) alcuni software fissano $b=0 \longrightarrow$ il vincoli lineare scompare. In alternativa si può scegliere di ottimizzare rispetto ad una nuova variabile:

$$\tilde{w} = \begin{pmatrix} w \\ b \end{pmatrix} \quad \tilde{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$$

$$w^T x + b \longleftrightarrow \tilde{w}^T \tilde{x}$$

La teoria di Vapnick cade in entrambi i casi, ma il duale diventa:

$$\begin{cases} \min \frac{1}{2} \lambda^T Q \lambda - e^T \lambda \\ 0 \leqslant \lambda_i \leqslant C \\ i = 1, \ldots, l \end{cases}$$

che, considerando una variabil per veolta e introducendo il **metodo delle coordinate duali** $|W_k| = 1$, la soluzione del sottoproblema diventa analitica ed è possibile effettuare lo shrinking.

Nel caso di Kernel Lineare si può usare esplicitamente:

$$w = \sum_{i=1}^{l} \lambda_i^* \, y^i \, x^i$$

come legge di aggiornamento:

$$w^k = \sum_{i=1}^{l} \lambda_i^k \, y^i \, x^i$$

che al limite converge a w^*.

Il calcolo del gradiente migliora diventanto $O(n) \ll O(l)$ del costo di una singola iterazione.

19 Problemi di Regressione con SVM

Le SVM possono essere utilizzati per **problemi di regressione**:

$$\text{TS} = \{(x^i, y^i), x^i \in \mathbb{R}^n, y^i \in \mathbb{R} \quad i = 1, \ldots, l\}$$

Si impone che l'errore:

$$|y^i - w^T x^i - b| \leqslant \varepsilon$$

sia minore id una certa tolleranze ε.

Il vincolo si trasforma in una coppia di vincoli lineari del tipo:

$$y^i - w^T x^i - b \leqslant \varepsilon - \xi_i$$

$$y^i - w^T x^i - b \geqslant -\varepsilon + \hat{\xi}_i$$

$$\xi_i \geqslant 0 \quad \hat{\xi}_i \geqslant 0$$

19.1 Regressione Logistica

La **regressione logistica** è uno strumento di classificazione lineare che utilizza il concetto di distanza dell'iperpiano. In particolare, si fornisce una misura dell'importanza delle features:

$$\text{TS} = \{(x^i, y^i), x^i \in \mathbb{R}^n, y^i \in \mathbb{R} \quad i = 1, \ldots, l\}$$

Si modella direttamente le probabilità di appartenenza ad una classe.

Si definisce:

$$\Theta \in \mathbb{R}^{n+1} \quad \theta^{\circ} := \text{offset} \quad \theta_i = \text{peso della feature } i$$

La probabiità viene modellata tramite la dunzione logistica:

Figura 7.

che approsisma la funzione gradino. Quindi:

$$P(C=1|x_1=\bar{x}_1,\ldots,x_n=\overline{x_n}) = \frac{1}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}} = \cdot\frac{e^{(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}}$$

In cui $(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})$ è un iperpiano e la distanza di un punto x è proporzionale al suo modulo.

Nota 109.

$$P(C=1|x_1=\bar{x}_1,\ldots,x_n=\overline{x_n}) + P(C=0|x_1=\bar{x}_1,\ldots,x_n=\overline{x_n}) = \frac{1}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}} = \cdot\frac{e^{(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}} + \frac{1}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}} = 1$$

Corrisponde a considerare la distanze dei punti del Training Set dell'iperpiano. Maggiore probabilità fornisce una maggiore confidenza della classificazione,

19.2 Come si addresta la regressione logistica?

Per addrestrare una funzione di errore di masismizzazione si prende una **funzione di massima verosimiglianza**:

$$L(\Theta) = \prod_{x^k\in \mathrm{TS}^+} \frac{1}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}} \prod_{x^k\in \mathrm{TS}^-} \frac{1}{1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}}$$

Data la compelssità delle produttorie si massimizza $L(\Theta)$ a cui si applica il logaritmo:

$$\max_{\Theta} LL(\Theta) = -\sum_{x^k\in \mathrm{TS}^+} \log(1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})}) - \sum_{x^k\in \mathrm{TS}^-} \log(1+e^{-(\theta_o+\sum_{i=1}^{n}\theta_i\overline{x_i})})$$

Questo è un problema di **massimizzazione non vincolata** e $LL(\Theta)$ è concava in Θ, θ. Allora tutti e soli i punti che annullano il gradiente sono ottimi globali.

Importante 110. $\nabla LL(\Theta)^T d \geqslant 0 \forall d$ ammissibile $\longleftrightarrow \nabla LL(\Theta) = 0$ per qualsiasi d ammissibile

Nota 111. Nel caso in cui si abbia $\min_{x\in\mathbb{R}^n} f(x)$ quello che voliamo è una direzione d di discesa se e solo se $\nabla f(x)^T d < 0 \longrightarrow d - \nabla f(x)$ se ho $\max f(x)$.

In particolare:

$$\frac{\partial \mathrm{LL}}{\partial \theta_i} = \sum_{x^k\in \mathrm{TS}^+} \frac{e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})} x_i^k}{1+e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})}} - \sum_{x^k\in \mathrm{TS}^-} \frac{e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})} x_i^k}{1+e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})}} =$$

$$= \sum_{x^k\in \mathrm{TS}^+} \frac{x_i^k}{1+e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})}} - \sum_{x^k\in \mathrm{TS}^-} \frac{x_i^k}{1+e^{-(\theta_o+\sum_{i=1}^n \theta_i \overline{x_i})}}$$

In pratica si stanno utilizzando le probabilità di erroe per effettuare l'addestramento:

$$= \sum_{x^k\in \mathrm{TS}^+} x_i^k \Pr(x_k \in \mathrm{TS}^-) - \sum_{x^k\in \mathrm{TS}^-} x_i^k \Pr(x^k \in \mathrm{TS}^+)$$

Inoltre si pone:

$$\theta_i^k \to \theta_i^k + \alpha^k \left(\sum_{x^k\in \mathrm{TS}^+} x_i^k \Pr(x_k \in \mathrm{TS}^-) - \sum_{x^k\in \mathrm{TS}^-} x_i^k \Pr(x^k \in \mathrm{TS}^+) \right)$$

Importante 112.

Il paramero α_k deve essere opportunamente scelto.

$|\vartheta_i|$ è una misura dell'importanza della feauture $i-$esima. Per ridurre l'overfitting si aggiunge un termine di regolaizzazione scelto in cross validation e strettamente positivo:

$$\max \mathrm{LL}(\Theta) - \alpha \frac{1}{2} \|\Theta\|^2$$

19.3 Valutazione di un classificatore in probabiltà

Per valutare un classificatore in probabilità, si passa da un assegnamento soft ad uno har scegliente una soglia t:

$$P(C=c|x) > t \longrightarrow C = c$$

Per ogni valore della soglia t posso definre:

$$S(t) := \text{insieme ottenuto con la soglia } t \text{ di punti nella classi} + 1$$

Un'altra tecnica è quella del **G grand truth**(veri positivi) in cui si calcola al variare di t la precision:

$$\mathrm{Precision}(t) = 100 \frac{|S(t) \cap G|}{|S(t)|} := \text{percentuale di positivi dichiarati veramente positivi}$$

Importante 113. Non è detto che all'aumentare di t la precision sia monotona

Inoltre, posso devinre la recall:

$$\mathrm{Recall}(t) = 100 \frac{|S(t) \cap G|}{|G|} := \text{percentuale di positivi effettivamente giusti}$$

A questo punto posso calcolare la **F1-Score**, è una media armonica che prende in considerazione la recall e la precision:

$$F1(t) = \frac{2\,\text{Precision}(t)\,\text{Recall}(t)}{\text{Precision}(t) + \text{Recall}(t)}$$

19.4 ROC CURVE (receiver operating characteristic)

Per confrontare due classificatori posso utilizzare la ROC CURVE

$$\text{TPR}(t) = \text{Recalla}(t) = 100\,\frac{|S(t) \cap G|}{|G|} \quad \text{oridnata}$$

$$\text{FPR}(t) = \frac{|S(t) - G|}{|\text{TS} - G|} \quad \text{ascissa}$$

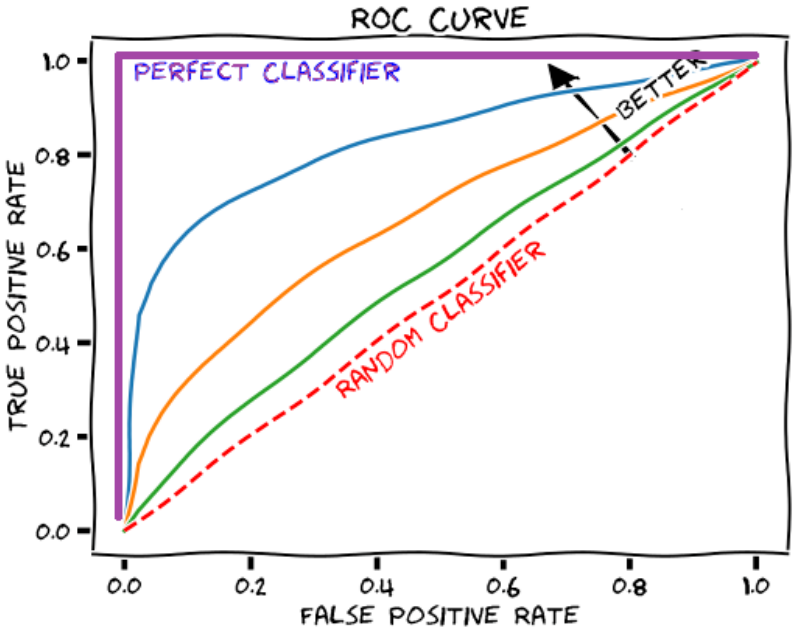


Figura 8.

Definiamo:

- **AUC:**area sotto la curva e tanto più è alta tanto è più buono il classificatore. Questo è un parametro indipendente da t.

20 Clustering

Il clustering è un tipo di classificazione **non supervisionata**, cioè ad ogni punto del Training Set non sono note ne le effettive uscite ne quante uscite ci si può aspettare.

Dato un Training Set $\text{TS} = \{x^i \in \mathbb{R}^n \quad i = 1, \ldots, N\}$ vogliamo raggruppare i punti del TS molto simili in **clust** e in clust diversi punti non simili. In particolare, s vuole partizionare l'insime di M cluster $C_1, \ldots, C_M$ in modo tale che:

$$C_i \cap C_j = \varnothing \quad e \sum_{i=1}^{M} C_i = \text{TS}$$

A tale scopo, si necessita una definizione di dissimilità tra punti e per arrivare a questa è necessario avere une nozione di distanza tra attibuti:

1. **Attributi Continui** ($x_{ij} \in \mathbb{R}$): si può usare $l(|x_{ij} - x_{kj}|) \quad i,\ k = 1, \ldots, N$ l monotona crescente (es.norma);

2. **Attibuti Ordinali:** si asusmono k valori tra loro ordinati e ci si riconduce a variabili quantitative traducendo k valore numeri che rispettano l'ordinamento originale.

3. **Attributi Categorici Non Ordinati:** si deve definire la distanza ad-hoc: se ho de valori r e r':$L_{rr} = 0, L_{rr'} \geqslant 0$

Una volta definita la distanza tra due attributo:

$$l_j(x^i, x^k) := \text{distanza da attributi } j$$

$$d(x^i, x^k) = \sum_{j=1}^{n} w_j\, l_j(x^i, x^k) \quad w_j = 1 \quad j = 1, \ldots, n$$

Importante 114. I dati non venogno scalati poiché essendo non supervisionato non si hanno uscite

Definizione 115. *Variabili di Assegnamento*

$$\delta_{ij} = \begin{cases} 1 & \text{se } i \text{ è assegnato al cluste } j \quad i = 1, \ldots, N \\ 0 & \text{altrimenti} \qquad\qquad\qquad\quad j = 1, \ldots, M \end{cases}$$

Definizione 116. *Variabile del centroide*

$z_j \in \mathbb{R}^n$ ***centroide del cluster j*** *(è un punto significativo del cluste)* $j = 1, \ldots, M$

Per misurare la bontà di un cluster si utilizza la **distanza intraclasse**:

$$s_j = \sum_{i=1}^{N} \delta_{ij}\, d(x^i, z_j)$$

Il problema di clustering risulterà, quindi, del tipo:

$$\begin{cases} \min_{\delta, z} \sum_{j=1}^{M} \sum_{i=1}^{N} \delta_{ij}\, d(x^i, z_j) \\ \sum_{j=1}^{M} \delta_{ij} = 1 \quad i = 1, \ldots, N \text{ (assegamento del punto ad un solo cluster)} \\ \delta_{ij} \in \{0, 1\} \quad z_j \in \mathbb{R}^n \quad i = 1, \ldots, Nj = 1, \ldots, M \end{cases}$$

Prendendo in considerazione come distanza, la distanzia euclidea:

$$d(x^i, z_j) = \|x^i - z_j\|^2$$

Posiamo scrivere il problema primale:

$$\begin{cases} \min_{\delta, z} \sum_{j=1}^{M} \sum_{i=1}^{N} \delta_{ij}\, \|x^i - z_j\|^2 \\ \sum_{j=1}^{M} \delta_{ij} = 1 \quad i = 1, \ldots, N \text{ (assegamento del punto ad un solo cluster)} \\ \delta_{ij} \in \{0, 1\} \quad z_j \in \mathbb{R}^n \quad i = 1, \ldots, Nj = 1, \ldots, M \end{cases}$$

Che può essere scritto in una forma equivalente:

$$\begin{cases} \min_{\delta,z} \sum_{j=1}^{M} \sum_{i=1}^{N} \delta_{ij} \|x^i - z_j\|^2 \\ \sum_{j=1}^{M} \delta_{ij} = 1 \quad i = 1, \ldots, N \text{ (assegamento del punto ad un solo cluster)} \\ \delta_{ij} \geqslant 0 \quad z_j \in \mathbb{R}^n \quad i = 1, \ldots, N j = 1, \ldots, M \end{cases}$$

Per mostrare che i due problemi sono equivalenti deve esistere una trasformazione tra i punti ammissibili di un problema all'altro e viceversa.

Teorema 117. *Teorema Equivalenza P e Q*

Dati:

$$\begin{cases} \min f(x) \quad (P) \\ x \in S \end{cases} \qquad \begin{cases} \min g(y) \quad (Q) \\ y \in D \end{cases}$$

Se:

1. $\forall x \in S. \exists y \in D \colon g(y) \leqslant f(x) \longrightarrow \forall x \in S, \exists \rho(x) \in D \colon g(\rho(x)) \leqslant f(x)$
2. $\forall y \in D, \exists x \in S \colon f(x) \leqslant g(y) \longrightarrow \forall y \in D, \exists \sigma(y) \in S \colon f(\sigma(y)) \leqslant g(y)$

Allora il problema (P) ammette minimo globale se e solo se (Q) ammette minimo globale.

Dimostrazione.

Tenendo in considerazione il problema (Q), provando a fissare una delle due varibiali:

- $\delta = \overline{\delta}$:

$$\min_{z} \sum_{j=1}^{M} \sum_{i=1}^{N} \overline{\delta_{ij}} \|x^i - z_j\|^2$$

Importante 118. I vincoli si levano

Lo si può decomporre in M problemi indipendenti:

$$\min_{z_j} \sum_{i=1}^{N} \overline{\delta_{ij}} \|x^i - z_j\|^2$$

Questo problema è coercivo e convesso in assenza di vincoli, quindi è possibile trovare il minimo globale annullando il gradiente rispetto a z:

$$\nabla_{z_j} f(\overline{\delta}, z) = 2 \sum_{i=1}^{N} \overline{\delta_{ij}} (x^i - z_j) = 0$$

$$\sum_{i=1}^{N} \overline{\delta_{ij}} x^i - \sum_{i=1}^{N} \overline{\delta_{ij}} z_j = 0$$

$$z_j = \frac{\sum_{i=1}^{N} \delta_{ij} x^i}{\sum_{i=1}^{N} \overline{\delta_{ij}}}$$

In cui, il denominatore corrisponde a $|C_j|$ e la variabile z_j rappresenta il **baricentro** dei punti nel cluster j ed è anche l'ottimo globale. Inoltre, data l'assegnazione $\overline{\delta_{ij}}$ binaria non si può fare di meglio.

- $z = \overline{z}$:

$$\min_{\delta} \sum_{j=1}^{M} \sum_{i=1}^{N} \delta_{ij} \, \|x^i - \overline{z_j}\|^2 = \sum_{i=1}^{N} \delta_j^T d_j \quad \textbf{Lineare}$$

$$\delta_j = \begin{pmatrix} \delta_{1,j} \\ \vdots \\ \delta_{1.M} \end{pmatrix} \quad d_j = \begin{pmatrix} \|x^1 - \overline{z_j}\|^2 \\ \vdots \\ \|x^N - \overline{z_j}\|^2 \end{pmatrix}$$

$$s.t. \begin{cases} \sum_{j=1}^{N} \delta_{ij} = 1 \\ \delta_{ij} = 0 \end{cases}$$

è un problema di **simplesso** unitario.

Nota 119. Simplesso Unitario

Il problema è nella forma:

$$\begin{cases} \min f(x) \\ \sum_i x^i = 1 \\ x^i \geqslant 0 \end{cases}$$

Scriviamo le condizioni di KKT necessarie e sufficienti di otitmo globali introducento i moltiplicatori rispettivamente μ, λ_i. Notiamo preliminarmente che l'insieme è compatto, quindi la soluzione esiste.

La lagrangiana è:

$$L(x, \lambda, \mu) = f(x) + \mu\left(\sum_i x^i - 1\right) - \sum_{i=1}^{N} \lambda_i x^i$$

Annullamento della Lagrangiana:

$$\frac{\partial f}{\partial x_i}(x) + \mu - \lambda_i = 0 \qquad i = 1, \ldots, n$$

Complementarietà:

$$\lambda_i \, x_i = 0 \quad i = 1, \ldots, n$$

$$\lambda_i \geqslant 0$$

Ammissibilità:

$$\sum_{i=1}^{N} x_i = 1$$

$$x_i \geqslant 0$$

Supponiamo che $x_i > 0 \longrightarrow \lambda_i = 0$:

$$\frac{\partial f}{\partial x_i} + \mu = 0 \longrightarrow \mu = -\frac{\partial f}{\partial x_i}$$

Sostituita nelle altre condizioni, $\forall j \neq i$:

$$\frac{\partial f}{\partial x_j} - \frac{\partial f}{\partial x_i} - \lambda_j = 0 \qquad j = 1, \ldots, n \quad j \neq i$$

Allora:

$$\lambda_i = \frac{\partial f}{\partial x_j} - \frac{\partial f}{\partial x_i} \geqslant 0$$

Quindi, se $x_i > 0$:

$$\frac{\partial f}{\partial x_j} \leqslant \frac{\partial f}{\partial x_i} \quad \forall j = 1, \ldots, n$$

Che rappresenta la **condizioni di ottimo per problemi di simplesso**.

Ricapitolando nel nostor problema:

$$\frac{\partial f}{\partial \delta_{ij}} = \|x^i - \overline{z_j}\|^2$$

Se il punto viene assegnato al cluster:

$$\delta_{ij} > 0 \longrightarrow \frac{\partial f}{\partial \overline{\delta_{ij}}} \leqslant \frac{\partial f}{\partial \delta_{ij}} \longrightarrow \|x^i - \overline{z_{\bar{j}}}\|^2 \leqslant \|x^i - \overline{z_j}\|^2 \quad \forall j \neq \bar{j}$$

Si ottiene l'ottimo globale assegnando il punti i al centroide più vicino:

$$\delta^*_{ij} = \begin{cases} 1 \,\text{se}\, j = \bar{j} \\ 0 \,\text{altrimenti} \end{cases} \longrightarrow \text{Equivalenza soddisfatta} \qquad \square$$

Riassumendo, (Q) è un **rilassamento** di (P) e per andare da (P) a (Q) non necessitiamo di trasformazioni poiché la funzione obiettivo è la stessa. Al contrario, se dobbiamoa ndare da (Q) a (P):

$$\begin{pmatrix} \overline{\delta} \\ \overline{z} \end{pmatrix} \in Q \rightarrow \begin{pmatrix} \hat{\delta} \\ \overline{z} \end{pmatrix} \text{dove } \hat{\delta}_{ij} = \begin{cases} 1 & \text{se } d(x^i, \overline{z_j}) \leqslant d(x^i, \overline{z_k}) \quad \forall k \neq j \\ 0 & \text{altrimenti} \end{cases}$$

Con la proprietà che:

$$\begin{pmatrix} \hat{\delta} \\ \overline{z} \end{pmatrix} \in (P) \quad e \quad f(\hat{\delta}, \overline{z}) \leqslant f(\overline{\delta}, \overline{z})$$

Dato che $\hat{\delta}$ è l'ottimo globale, allora i due problemi sono equivalenti.

20.1 Algoritmo di K-Means

Algoritmo

Dati $z^\circ \in \mathbb{R}^n, k = 0$
While criterio di arresto soddisfatto

1. Assegmanento dei punti:
   For $i = 1, \ldots, N$ poni;

$$\delta^{k+1}_{ij} = \begin{cases} 1 & \text{se } j = j^* \text{ ovvero } \arg\min \|x^i - z_j\|^2 \\ 0 & \text{altrimenti} \end{cases}$$

2. Ricalcolo dei centroidi:
   For $j = 1, \ldots, M$ poni:

$$z^{k+1}_j = \frac{\sum \delta^{k+1}_{ij} x^i}{\sum_{i=1}^N \delta^{k+1}_{ij}}$$

End

Criterio di Arresto

$$\frac{f(\delta^{k+1}, z^{k+1}) - f(\delta^k, z^k)}{1 - f(\delta^k, z^k)} \leqslant \varepsilon$$

Teorema 120. *Teorema Convergenza K-Means*

L'algoritmo di K-Means con $\varepsilon = 0$ converge in un numero finito di passi in δ^*, z^* tale che:

$$f(\delta^*, z^*) \leqslant f(\delta . z^*) \quad \forall \delta \text{ ammissibile}$$

$$f(\delta^*, z) \leqslant f(\delta^*, z^*) \quad \forall z \in \mathbb{R}^m \text{ ammissibile}$$

Dimostrazione.

Supponiamo per assurdo che l'algoritmo non termini, quindi:

$$f(\delta^{k+1}, z^{k+1}) < f(\delta^k, z^k) \quad \forall k$$

Ad ogni passo:

$$f(\delta^k, z^{k+1}) \leqslant f(\delta^k, z) \quad \forall z \in \mathbb{R}^m$$

e:

$$f(\delta^{k+1}, z^k) \leqslant f(\delta, z^k) \quad \forall \delta \text{ ammissibile}$$

Allora $\exists$ sottosuccessione k tale che:

$$\{\delta^k, z^k\}_K \quad \delta^k = \overline{\delta}$$

$$\forall k \in K \, \exists \text{indice } l(k) \quad k + l(k) \in K$$

$$f(\delta^k, z^k) = f(\overline{\delta}, z^k)$$

$$f(\delta^{k+l(k)}, z^{k+l(k)}) = f(\overline{\delta}, z^{k+\text{ò}(k)})$$

Inoltre se:

$$f(\delta^k, z^k) = f(\overline{\delta}, z^k) \leqslant f(\overline{\delta}, z) \quad \forall z$$

Poiche $l(k) + k > k$ devo avere che:

$$f(\delta^{k+l(k)}, z^{k+l(k)}) < f(\delta^k, z^k) = f(\overline{\delta}, z^k) \leqslant f(\overline{\delta}, z = z^{k+l(k)})$$

Assurdo poiché abbiamo supposto che:

$$f(\overline{\delta}, z^{k+l(k)}) \leqslant f(\overline{\delta}, z = z^{k+l(k)})$$

□

20.2 Scelta del numero di Cluster

Dato che il problema di clusterizzazione è un problema di ottimizzazione non supervisionata, si ha difficoltà a scegliore il numero di cluster da utilizzare poiché non si conoscono il numero di uscite. Quindi ci sono delle euristiche che permettono di scegliere il numero di Cluster:

- **Elbow Method**:

Si definisce **elbow** il punto in cui f non decresce di molto e si hanno le sequenti proprietà:

$$f_k^* \ll f_k \text{ quando } k < k^*$$

$$f_{k+1}^* \approxeq f_k^* \quad \text{quando } k > k^*$$

- **Silhouette**:

  Si fissa k e si risolve per quel valore. In particolare:

  - $\forall i = 1, \ldots, N \quad \iota \in C(i)$, calcolo:
    - $a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i} d(x_i, x_j)$
    - $b(i) = \min_{k \neq C(i)} \frac{1}{|C_k|} \sum_{j \in K} d(x_i, x_j)$
    - $S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \text{se } |C_i| > 0 \quad o S(i) = 0 \quad \text{se } |C(i)| = 1$
  - Al variare di k calcolo:

$$S_k = \frac{1}{N} \sum_{i=1}^{N} S(i)$$

  Dato che voglio il valore massimo, scelgo il k più alto.

PACIFICI

1 Alberi decisionali per la classificazione

1 Introduzione

Apprendimento statistico

L'**apprendimento statistico** sono delle tipologie di apprendimento che può essere svolta sia in maniera automatica sia tramite ML (Machine Learning).

In parciolare si compone:

$$x^{(i)} \in \mathbb{R}^p \rightarrow \textbf{osservazione}(\text{punto})$$

$$y^{(i)} \in \mathbb{R} \rightarrow \textbf{risposta} \quad \text{con} \quad i = 1, \ldots, n$$

Definiamo il piano/spazio ottenuto dalle osservazione, **spazio delle feature**.

Il nostro obiettivo è quello di ottenere la funzione di uscita in funzione delle osserviazioni date. Questo processo viene detto **inferire**. Formalmente:

$$y = f(x) \quad \text{con } x \in \mathbb{R}^p$$

Riassumendo il nostro problema, noti:

- $(x^{(i)}, y^{(i)})$ con $i = 1, 2, \ldots, n$ osservazioni disponibili$\rightarrow$**Training set**;

- $x^{(i)} \in \mathbb{R}^p :=$ variabili di input indipendenti o anche dette **feature**;

Ottenere:

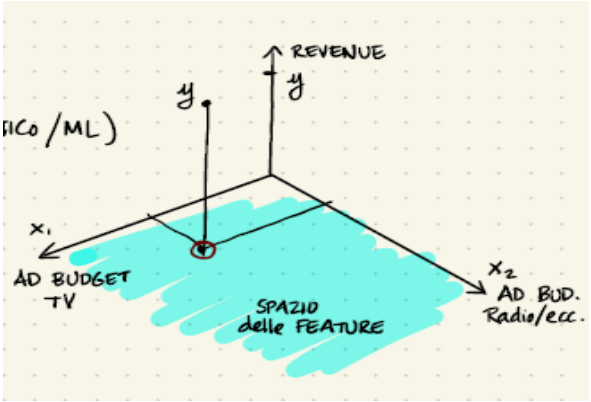
- $y \in \mathbb{R} :=$ variabile di output / **Class** / **etichetta** / **label**.

In altre parole, dato $x \neq x^{(i)}$ qual è la sua uscita? Per rispondere a questa domanda è necessario che la $f(x)$ resistuisca esattamente $y^i = f(x^{(i)})$, cioè che ci sia un **fitting** con i dati del **TS (Training Set)**.

Un modello di questo tipo è un **modello di inferenza/predittivi** poiché permettono di intuire la prossima uscita e producono una y data una certa x.

Apprendimento supervisionato

L'**apprendimento supervisionato** è un tipo di apprendimento per cui l'uscita y è associata un punto dello spazio delle feature. Graficamente:



Apprendimento non supervisionato

L'**apprendimento non supervisionato** è caratterizzato dal fatto che non esiste in corrispondendo di una osservazione nello spazio delle feature la relativa uscita.

Response

Il **response** è l'uscita y che si ottiene dal fitting con il TS.

La response può essere di due tiplogie:

- qualitativa: la **response** è discreta e finita e il suo risultato è una **classificazione** di punti in classi/label. Si cerca quindi di assegnare una classe ad una nuova osservazione conoscendo i dati a disposizione.

  Sono metodi basati su **alberi di decisione.**

- quantitativa: tipicamente è una uscita continua ed i metodi più utilizzati sono la Regressione, Super Vector Machine, Reti Neurali etc..

Quindi il nostre Data Set sarà composto dalle righe che corrispondono alle osservazioni, dalle colonne che sono gli attributi delle varie osservazioni ed a ogni riga, a cui corrispondere un'osservazone con i suoi attributi, è associrata una etichetta.

2 Alberi Decisionali

Gli alberi decisionali applicano ricorsivamente metodi euristici per creare delle divisioni nello spazio delle feature per mettere in evidenza un certo aspetto del dataset.

Per ottenere una lbero decisionale dato uno spazio delle feature contenente le osservazioni, occorre:

- **Partizionare/stratificare/segmentare** lo spazio delle feature in modo tale che le partizioni siano il più omogenee possibile in base alla prevalenza o meno di etichette uguali;
- Si associa a ciascuna regione un'etichetta che corrispondendo all'etichetta predominante nella regione anche in presenza di regioni miste;
- Ipotizzando che ci arrivi una nuova osservazione, essa cadrà nello spazio delle feature all'interno di una regione R_i (ottenuta dalle precedenti partizioni) ed assumerà l'etichetta della regione corrispondente;
- Si delimitano i confini delle regioni R_i delle osservazioni e vi si assegnano dei valori;
- L'albero che si formerà sarà costituito da **nodi foglia** e **nodi interni**:
    - i **nodi interni**, sono i **test** che ci permettono di classificare le osservazioni tramite la verifica degli attributi. In particolare se si esamina una sola feauture il test viene detto **univariato (|−)**; altrimenti **multivariato(/);**
    - i **nodi foglia** sono i risulati della nostra classificazione a cui viene associata l'etichetta della regione R_i *a* cui appartengono.

N.B.:

Confronto test univariato vs test multivariato:

- Nel caso di test univariato si potrebbero fare errori di classificazione, ma l'albero risultante è più semplice;
- Nel caso di test multivariato non vi sono errori di classificazione per etichette già note, ma è più soggetto ad errori rispetto al caso univariato(**errore di generalizzazione**), ma l'albero presenta una complessità maggiore data la presenza di numerosi nodi;

Gli attributi di un albero decisionale sono di due tipi:

- **categorici**: cioè qualitativi;
- **continui**: cioè quantitativi;

Dal **training data** cioè l'insieme degli attributi e dell'uscita, si ottiene l'**albero decisionale**.

Ovviamente l'albero risultante varia in base all'ordine in cui i test vengono considerati.

Ogni volta in cui si fa un test, si ha una divisione die punti del TS e il nostro obiettivo è quello di produrre degli insiemi più omogenei possibili su cui applicare il test set.

Pro vs Contro alberi di decisione

- Facilità da spiegare/interpretare e si mima il comportamento umani;
- Facile da manovrare e le variabili sono di tipo qualitativo;

– Strumenti meno accurati;

– "Nervosismo": piccole variazioni di dati portanto ad alberi differenti.

Per ovviare a questa ultima problematica si ricorre a dei **metodi di aggregazione** in cui si costruiscono diversi alberi per poi essere uniti tra loro.

TDITD Family

La **TDITD family** compende una serie di Top-Down Induction Decision Trees come il CART e ID3. Questi alberi utilizzando l'induzione, cioè la ricorsione, per creare degli split nello spazio delle feature. Tuttavia, questi potrebbero non rappresentare opportunamente le caratteristiche del dataset.

In particolare, l'albero è usato per classificare i punti del test set secondo gli split e le labels. Questi alberi vengono detti **top-down** poiché, partendndo dal nodo radice, determinano uno split risolvere un problema di ottimizzazione, in genere di minimo rispetto ad una misura di inpurità, per poi applicare la stessa regola alle foglie appena generate.

Questo tipo di classificazione, inoltre, viene detta **greedy**, poiché ogni divisione viene determinata in isolamento senza considerare i possibili impatti che si potrebbero avere nel furuo.

In conclusione, questa famiglia di alberi soffre problemi di complessità.

Algoritmo di Hunt

Algoritmo

1. **INIT:**
   list:={training set}
   node_list:={root}, Droot:=list: insieme delle proprietà che soddisfano i test
   test_pool:={...}
2. **Pick t ∈node list**
3. **Definire $D_t \subseteq$ list/*punti del TS che raggiungono il nodo t*/**
4. **IF $D_t = \varnothing$ THEN t è LEAF label t as "*" (null)**
5. **IF D_t ha elementi tutti di classe K THEN t è LEAF label t as k**
6. **IF D_t ha elementi di classi diverse THEN**
   a. **Choose test ∈test_pool**
   b. **split di D_t in sottoinsiemi $D_t^1, D_t^2, \ldots, D_t^p$**
   c. **update node_list:= node_list$\cup\{t_1\} \cup \{t_2\} \cup \cdots \cup \{t_p\}$**
7. **node_list := node_list$\setminus${t}**
8. **IF node_list=$\varnothing$ THEN STOP ELSE goto 1**

Algoritmo ID3

Algoritmo
ID3 (idea)
1. **Selezione una "window" (sottoinsieme del TS);**
2. **Costruisci un DT usando i punti della window;**
3. **IF ∃ elementi ∈ TS$\setminus${window} che non sono classificati correttamente THEN**
   window:=window ∪questi elementi
   ELSE Return DT and STOP

3 Errori negli alberi di classificazione

- **Errore di classificazione:**= sDato un DT, si ha errore di classificazione se il numero di elementi del TS che vengono etichettato in modo corretto. Una sua minimizzazione troppo marcata non porta ad un buon albero poiché soffre dell'**overfitting**. In aggiunta, nel caso di **overfitting**, si eliminano alcuni rami per avere un albero più "semplice", **Tree Pruning**;

- **Errore di generalizzazione:**= è un errore atterso che il DT compie sui punti $\notin$TS. Viene stimato con il test set.

Indici di (in)purity

Nella TDIDT "family" la costruzioni dei DT, di norma si sceglie la **strategia greedy**: in ogni passo si sceglie il test che ottimiza un dato criterio disinteressandosi completamente di quello che può accadere in seguito.

Tuttavia, vogliamo un certo livello di **omogeneità**. A tale scopo si introducono i **purity indices** che indica quanto un sottoinsieme è omogeneo, cioè composto dalle stesse classi.

Errore di classificazione

Sia $p_i \in [0,1]$ frequenza con cui l'etichetta i comprare nell'insieme. Allora:

$$e = 1 - \max_i \{p_i\} \longrightarrow \arg_i \max \{p_i\} := \text{moda}$$

$$0 \leqslant e \leqslant 1 - \frac{1}{k} \quad \text{con } k := \#\text{di etichette}$$

Gini Index

$$g = 1 - \sum_{i=1}^{k} p_i^2 \qquad 0 \leqslant g \leqslant 1 - \frac{1}{k}$$

Entropia

$$E = -\sum_{i=1}^{k} p_i \log_k(p_i) \qquad 0 \leqslant E \leqslant 1$$

Dimostrazione.

Si ha omogeneità massima $p_i = 1, p_j = 0 \qquad \forall j \neq i$

Per $p = \frac{1}{k} \quad \forall i$, si ha che:

$$E = -\sum_{i=1}^{k} \left(\frac{1}{k}\right) \log_k\left(\frac{1}{k}\right) =$$

$$= \sum_{i=1}^{k} \left(\frac{1}{k}\right) (-1) \log_k\left(\frac{1}{k}\right) =$$

$$= \sum_{i=1}^{k} \left(\frac{1}{k}\right) \log_k\left(\frac{1}{k}\right)^{-1} = \sum_{i=1}^{k} \left(\frac{1}{k}\right) 1 = 1$$

□

Importante 121. Noi vogliamo usare gli indici per capire il tipo di test da utilizzare per generare foglie con un grado di omogeneità maggiore. A tale scopo, si sommano gli indici di impurità dei figli:

Test

Si possono effettuare vari tipi di test, i più comuni sono i seguenti:

- **ERR_TEST**=$\sum_{j \quad \text{figlio}}$ ERR_CLASS(j); si sceglie il test con indice minonre
- **GINI_TEST**=$\sum_{j\,\text{figlio}}$ GINI(j) $\frac{n_j}{n}$ si sceglie il test con indice minore

  In cui n_j sono i punti nel nodo figlio e n i punti nel nodo padre
- **INFOMATION GAIN_TEST**= misura quando si guadagna in omogeneità negli insieme con un certo tipo di test:

$$\text{INFORMATION\_GAIN\_TEST} = \text{ENTROPY(padre)} - \sum_{j \quad \text{figlio}} \left(\frac{n_j}{n}\right) \text{ENTROPY}(j)$$

  Si sceglie il test con indice maggiore;
- **GAIN_RATIO_TEST**= equivale all'information gain test, ma con un andamento uniforme:

$$\text{GAIN\_RATIO\_TEST} = \frac{\text{INFORMATION\_GAIN\_TEST}}{\left[-\sum_{j\,\text{figlio}} \left(\frac{n_j}{n}\right) \log\left(\frac{n_j}{n}\right)\right]}$$

4 Albero Ottimo di Classificazione

Introduzione

Il problema di albero decisionale ottimale tenta di risolvere il problema di classificazione di un dataset creando un albero decisione per ottenere l'ottimalità globale.

Assumiamo:

- Sia dato un Training Data (X,Y) contenente n osservazioni $(x_i, y_i), i = 1, \ldots, n$ ognuna delle quali con p feature;
- Modello univariato, cioè i test sono effettuati su un attributo alla volta;
- $y_i \in \{1, \ldots, k\}$ K classi/etichette;
- Attributi/feature continue (quantitative) $x_i \in [0,1]^p$;
- Criterio: min(errore di classificazione + complexity).

I metodi degli alberi decisionali tentano di effettuare una partizione ricorsiva $[0,1]^p$ per ottenere delle regioni disgiunte che rappresentano l'albero decisionale. L'albero finale sarà formato da nodi foglia e nodi di diramazione:

- I **nodi di diramazione** si dividono con parametri $\boldsymbol{a}$ e $\boldsymbol{b}$. Per nu dato punto i, $a^T x_i < b_i$ il punto seguirà la diramazione sinistra dal noto, altrminenti la destra;

- I **nodi foglia** sono assegnati ad una classe che determinerà la predizione di tutti i punti che cadono all'interno della folgia. La classe viene assegnata in base alla moda della foglia.

Decisioni del problema

Nella creazion dell'albero, in ogni iterazione, dobbiamo effettuare un numero fissato di decisioni:

- In ogni nodo dobbiamo decidere se dividere o fermaci;
- Dopo che abbiamo scelto di fermarci in un nodo, dobbiamo scegliere l'etichetta da assegnare al nuovo nodo foglia;
- Dopo che abbiamo scelto di dividere, dobbiamo scegliere su quali delle variabili effettuare il test;
- Quanndo classifichiamo i punti di traning secondo la costruzione dell'albero, dobbiamo decidere a quali dei nodi foglia assegnare un punto tale che la struttura dell'albero viene rispettata.

Struttura del problema

Consideriamo il problema di costruire un albero ottimale di decisione con la massimo profondità di D. Dato questo parametro:

- Albero binario completo (sotto albero di);
- Profondità fiddasa D$\Longrightarrow$# nodi $2^{D+1}-1=\mathbb{T}$

Notazione 122.

- $t \in \{T_B \cup T_L\} \setminus \{1\}$;
- $p(t) :=$ *nodo padre/genitore;*
- $A(t)$ *è l'****insieme degli antenatii*** *del nodo t;*
- $A_L(t), A_R(t)$ *sono rispettivamente gli* ***insiemi antenati di sinistra e di destra****, presi, partendo dalla radice fino a t, scegliendo rispettivamente la direzione destra o sinistra.;*
- $A(t) = A_L(t) \cup A_R(t)$

Quindi dividiamo i nodi in:

- **Branch Nodes:** sono tutti i nodi $t \in T_B = \left\{1, 2, \ldots, \left\lfloor \frac{T}{2} \right\rfloor\right\}$ che applicano la diisione nella forma $a^T x < b$ se lo soddisfano t seguirà il ramo sinistro, altrimenti il desto.

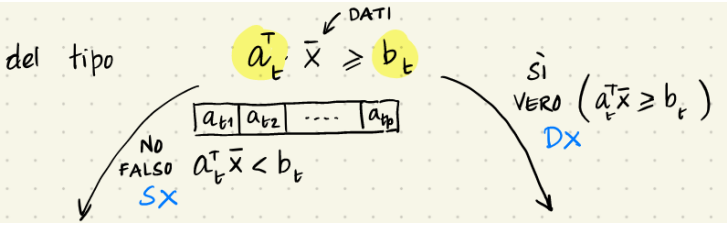


Figura 9.

- **Leaf Nodes**: Sono tutti i nodi $t \in T_L = \left\{ \left\lceil \frac{T}{2} \right\rceil, \ldots, T \right\}$ formano una classe di predizione per ogni punti che cade nel nodo foglia.

Variabili

Supponiamo $t \in T_B$

- $a_t \in \mathbb{R}^p$ coefficiente del test $\longrightarrow a_t \in \{0,1\}^p \quad \overline{x} \in [0,1]^p \qquad \sum_{i=1}^{p} a_t = 1$

$$(a_t \cdot \overline{x}) \in [0,1]$$

- $b_t \in \mathbb{R} \longrightarrow b_t \in [0,1]$ è uno scalare
- $d_t \in \{0,1\}$ variabile che indica se si effettua o no un test $d_t \in \{0,1\}$

Nota 123.

L'idea è quella di non permettere la divisione in un branch node. A tale scopo utilizziamo la variabile d_t per sapere su quale branch node si applica la divisione. Se un nodo di diramazione non applica lo split, allora viene modellato con $a_t = 0$ e $b_t = 0$.

Di conseguenza, si forzano tutti a seguire il ramo destro del nodo t. Questo permette all'albero non smettere di crescere.

Vincoli

1. $\sum_{i=1}^{p} a_{t,i} = d_t \qquad \forall t \in T_B \quad$ se effettuo o no il test
2. $0 \leqslant b_t \leqslant d_t \qquad \forall t \in T_B$
3. $d_t \leqslant d_{p(t)} \qquad \forall t \in T_B \quad$ rappresenta la coerenza con d_t se non si effetua lo SPLIT no split.

Allocazione

Il problema di **allocazione** riguarda l'assegnazione dei punti del training set alle foglie dell'albero decisionale e associarli agli errori che sono indotti da questa struttura.

Variabili

- Indichiamo con $z_{it} = \{0,1\} \quad i = 1, \ldots, n, t \in T_L$ se il punto x_i appartiene o meglio alla foglia t;
- Per verificare se una foglia possiede dei punti. Utilizziamo:

$$l_t \in \{0,1\} \quad t \in T_L \quad l_t = \begin{cases} 1 & \text{se la foglia } t \text{ ha assegnati dei punti} \\ 0 & \text{altrimenti} \end{cases}$$

- Il numero minimo di punti in ogni foglia $N_{\min}$,

Vincoli

$$\begin{aligned} z_{it} &\in \{0,1\} \quad i = 1, \ldots, n \quad t \in T_L \\ \sum_{t \in T_L} z_{it} &= 1 \quad i = 1, \ldots, n \quad \text{deve } \exists \text{una foglia per il punto } i \\ \sum_{i=1}^{n} z_{it} &\geqslant N_{\min} l_t \quad t \in T_L \quad \text{Una foglia deve essere sufficientemnte popolata} \end{aligned}$$

A questo punto dobbiamo aggiungere dei vincoli di coerenza che impongono le divisioni richieste fino al raggiungimento della foglia:

- Per i test **non soddisfatti**, deve valere: (9)

$$a_m^T x_i < b_i + M_1(1 - z_{it}) \quad i = 1, \dots, n \quad \forall t \in T_B, \forall m \in A_L(t)$$

- Per i test **soddisfatti**, deve valere:

$$a_m^T x_i \geqslant b_i - M_2(1 - z_{it}) \quad i = 1, \dots, n \quad \forall t \in T_B, \forall m \in A_R(t)$$

Nota 124.

Il vincolo •(9) non è supportato dei risolutori. Quindi occorre convertirlo in una forma che non utilizza la disuguaglianza stratta. A tale scopo, aggiungiamo un fattore piccolo ε e il vincolo diventa:

$$a_m^T x_i + \varepsilon \leqslant b_i + M_1(1 - z_{it}) \quad i = 1, \dots, n \quad \forall t \in T_B, \forall m \in A_L(t)$$

Tuttavia, se ε è troppo piccolo, potrebbe casuare **instabilità** numeriche, quindi occorre renderlo il più grande possibile senza influenzare la soluzione in qualsiasi problema valido.

Occorre definire un e_j per ogni feature j; il più gradnde valore valido è la distanza diversa da 0 più piccola tra i valori adiacenti della feature j:

$$\varepsilon_j = \min \left\{ \left| x_j^{(i+1)} - x_j^i \right| \quad x_j^{(i+1)} \neq x_j^i \quad i = 1, \dots, n-1 \right\}$$

In cui x_j^i è il più grande valore della feature j-esima.

Dall nota 124 possiamo utilizzare ε nel vincolo, dove ε_j corrisponde alla feature sul quale effettuiamo lo split:

$$a_m^T(x_i + \varepsilon) \leqslant b_m + M_1(1 - z_{it}) \quad i = 1, \dots, n \quad \forall t \in T_B, \forall m \in A_L(t)$$

Scelta M_1 e M_2

M_1 e M_2 sono dette costanti big-M. Valutando i vincoli, sappiamo che:

- Il masismo valore di $a_m^T(x_i + \varepsilon) - b_m$ è $1 + \varepsilon_{\max}$ in cui $\varepsilon_{\max} = \max_j \{\varepsilon_j\}$. Quindi vogliamo:

$$M_1 \geqslant 1 + \varepsilon_{\max}$$

- Il massimo valore di $b_t - a_t^T x_i$ è 1 e quindi possiamo scegliere:

$$M_2 \geqslant 1$$

Ottenendo, in conclusione gli ultimi due vincoli:

$$a_m^T x_i + \varepsilon \leqslant b_i + (1 + \varepsilon_{\max})(1 - z_{it}) \quad i = 1, \dots, n \quad \forall t \in T_B, \forall m \in A_L(t)$$

$$a_m^T x_i \geqslant b_i - (1 - z_{it})$$

Funzione Obiettivo

La funzione obiettivo che vogliamo determinare si basa su due criteri di minimizzazione riguardo:

- la **complessità dell'albero**: corrisponde alla dimensione dell'albero decisionale ed è fissata a priori da D;
- **Errore di classificazione:** assegnamo ad un'etichetta non corretta il costo di 1 e ad una corretta il costo 0;

Dati

Sia:

$$y_{\text{ik}} = \begin{cases} +1 & \text{se } y_i = k \quad \text{il punto } i \text{ ha etichetta } k \\ -1 & \text{altrimenti} \end{cases} \qquad i = 1, \ldots, n; k = 1, \ldots, K$$

Inoltre, $N_{kt} :=$ il numero di punti per l'etichetta k nel nodo t e $N_t :=$ il numero di punti nel nodo t:

$$N_t = \sum_{i=1}^{n} z_{it} \quad t \in T_L$$

$$N_{kt} = \frac{1}{2} \sum_{i=1}^{n} (1 + y_{ik})\, z_{ik} \qquad k = 1, \ldots, k$$

Task - obiettivo

Vogliamo assegnare alle foglioe $t \in T_L$ con l_t=1 l'etichetta a lui associata. Essa è la mode delle etichette assegnate nel nodo t.

Inoltre dobbiamo assegnare un'etichetta ad ogni nodo t nell'albero, definita come:

$$c_t \in \{1, \ldots, K\}$$

Ovviamente l'etichetta ottima è quella che corrisponde alla moda di in quel nodo, cioè alla frequenza in cui quella determinata etichetta compare nel nodo:

$$c_t = \arg \max_{k=1,\ldots,K} \{N_{\text{kt}}\}$$

Al finte di tenere traccia della predizione di ogni nodo utilizziamo la variabile binaria c_{kt}:

$$c_{kt} = \begin{cases} 1 & \text{se } l'\text{etichetta della foglia } t \text{ è } k \\ 0 & \text{altrimenti} \end{cases}$$

A questo punto ci dobbiamo assicurare che in ogni nodo ci sia una solo predizione. A tale scopo introduciamo il vincolo:

$$\sum_{k=1}^{K} c_{kt} = l_t \quad \forall t \in T_L$$

Dato che sappiamo come scegliere l'etichetta in maniera ottimale in ogni nodo, definiamo L_t come l'**errore di classificazione** in ogni nodo che risulta uguale al numero di punti nel nodo minore del numero di punti dell'etichetta pù comune:

$$L_t = N_t - \max_{k=1,\ldots,K} \{N_{kt}\} = \min_{k=1,\ldots,K} \{N_t - N_{kt}\} \quad t \in T_L$$

Che può essere linearizzato nel seguente modo:

$$\begin{aligned} L_t &\geqslant N_t - N_{kt} - M\,(1 - c_{kt}) \quad k = 1, \ldots, K \quad \forall t \in T_L \\ L_t &\leqslant N_t - N_{kt} + M\,c_{kt} \qquad\quad k = 1, \ldots, K \quad \forall t \in T_L \\ L_t &\geqslant 0 \qquad \forall t \in T_L \end{aligned}$$

Dove M è una costante sufficientemente grande che rende il vincolo intattivo dipendendo dal valore di c_{kt}. Un possibile valore che possiamo prendere è $M = n$.

Il costo totale dell'errore di classificazione è dato da $\sum_{t \in T} L_t$ e la complessità è il numero di split nelll'albero, dato da $\sum_{t \in T_b} d_t$. A questo punto possiamo normalizzare l'errore di classificazione con una **soglia di accuratezza** $\hat{L}$, ottenuta dalla semplice predizione dell'etichette più popolari dell'intero dataset rendendolo di conseguenza, indipendente dalla quantità α.

Finalmente possiamo dunque scrivere la funzione obiettivo:

$$\min \frac{1}{\hat{L}} \sum_{t \in T_L} L_t + \alpha \sum_{t \in T_B} d_t$$

Riepilogo

Mettendo insieme tutte queste formulazioni delle sezioni precedenti possiamo scrivere il nostro modello OCT:

$$\min \frac{1}{\hat{L}} \sum_{t \in T_L} L_t + \alpha \sum_{t \in T_B} d_t$$

$$\begin{aligned} L_t &\geqslant N_t - N_{kt} - M\,(1 - c_{kt}) \quad k = 1, \ldots, K \quad \forall t \in T_L \\ L_t &\leqslant N_t - N_{kt} + M\,c_{kt} \qquad\quad k = 1, \ldots, K \quad \forall t \in T_L \\ L_t &\geqslant 0 \qquad \forall t \in T_L \\ N_{kt} &= \frac{1}{2} \sum_{i=1}^{n} (1 + y_{ik})\, z_{ik} \qquad k = 1, \ldots, k \\ N_t &= \sum_{i=1}^{n} z_{it} \qquad \forall t \in T_L \\ \sum_{k=1}^{K} c_{kt} &= l_t \quad \forall t \in T_L \\ a_m^T x_i &\geqslant b_i - M_2(1 - z_{it}) \quad i = 1, \ldots, n \quad \forall t \in T_B, \forall m \in A_R(t) \\ a_m^T x_i + \varepsilon &\leqslant b_i + (1 + \varepsilon_{\max})(1 - z_{it}) \quad i = 1, \ldots, n \quad \forall t \in T_B, \forall m \in A_L(t) \\ \sum_{t \in T_L} z_{it} &= 1 \quad i = 1, \ldots, n \\ \sum_{i=1}^{n} z_{it} &\leqslant l_t \quad t \in T_L \\ \sum_{i=1}^{n} z_{it} &\geqslant N_{\min}\, l_t \quad t \in T_L \\ \sum_{i=1}^{p} a_{t,i} &= d_t \qquad \forall t \in T_B \\ 0 \leqslant b_t &\leqslant d_t \qquad\quad \forall t \in T_B \\ d_t &\leqslant d_{p(t)} \qquad\quad \forall t \in T_B \\ z_{it}, l_t &\in \{0, 1\} \quad i = 1, \ldots, n \quad \forall t \in T_L \\ a_{jt}, d_t &\in \{0, 1\} \quad j = 1, \ldots, p \quad \forall t \in T_B \end{aligned}$$

Importante 125.

La difficoltà del modello è determinata dal numero di variabili z_{it} in quale sono $n \cdot 2^D$.

Inoltre, occorre specificare a priori tre parametri: la massima profondità dell'albero D, la dimensione minima della foglia $N_{\min}$ e il parametro di complessità. Questa scelta viene detta **tuning** (sincronizzazione).

Il tempo di calcolo di un OCT è nell'ordine dei minuti.

Warm Starts

I risolutori traggono grande beneficio quando gli viene fornito una soluzione intera ammissibili come un **warm start** per il processo di soluzione.

Inserendo una forte soluzone di **warm start** il risolutore incrementa notevolemnte la velocità con cui è in grado di generare soluzioni fortemente ammissibili. Quindi fornisce un upper-bound iniziale sulla soluzione ottima che permette di effettuare l'azione di **pruning** e inoltre fornisce un punto di partenza per la ricerca locale euristica.

Se noi avessimo una soluzione generata per la profondità D, questa soluzione è un valido **warm start** per la profondità $D+1$. Questo è importante poiché la difficoltà del problema aumenta con il crescere della profondità e quindi potrebbe risultare svantaggioro eseguire un problema MIO (Mixed-Integer-Optimization) con una piccola profondità per generare un forte warm start.

Quinsi, data una soluzione iniziale ammissibile è possibile confrontare la velocità del solutore nel determinare una sluzine ottima.

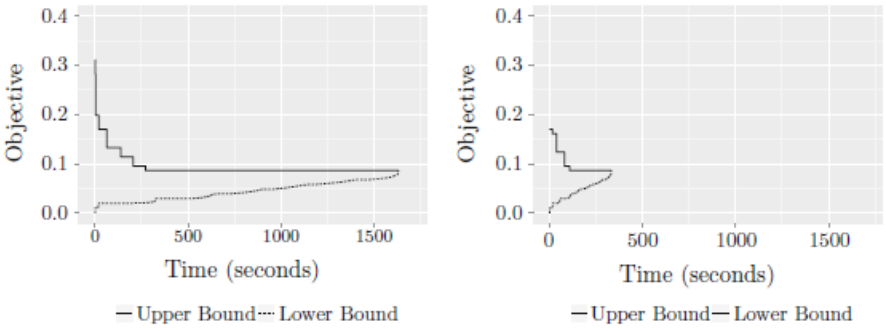


Figura 10. A sinistra senza Warm star; a destra con

Comi si può evincere dall'immagine, si ha convergenza quando upperbound e lower bound coincidono.

Tipicamente il warm start viene fornito da un algoritmo diverso.

5 Addestramento di un OCT

Addestrare una OCT consiste nello scegliere i parametri adatti al modello che vogliamo classificare. Questa azione viene detta **tuning** dei parametri e coinvolge la scelta di $N_{\min}, D, \alpha$ per minimizzare l'**erorre di training** e la **complessità dell'albero**

$$\min \frac{1}{\hat{L}} \sum_{t \in T_L} L_t + \alpha \sum_{t \in T_B} d_t$$

Importante 126.

- $\alpha :=$ è il fattore di importanze se è $=0 \longrightarrow$overfitting

Nota 127.

Un tipico approccio per la scelta di α è quello di discretizzare lo spazio di ricerca e i test di ogni valore.

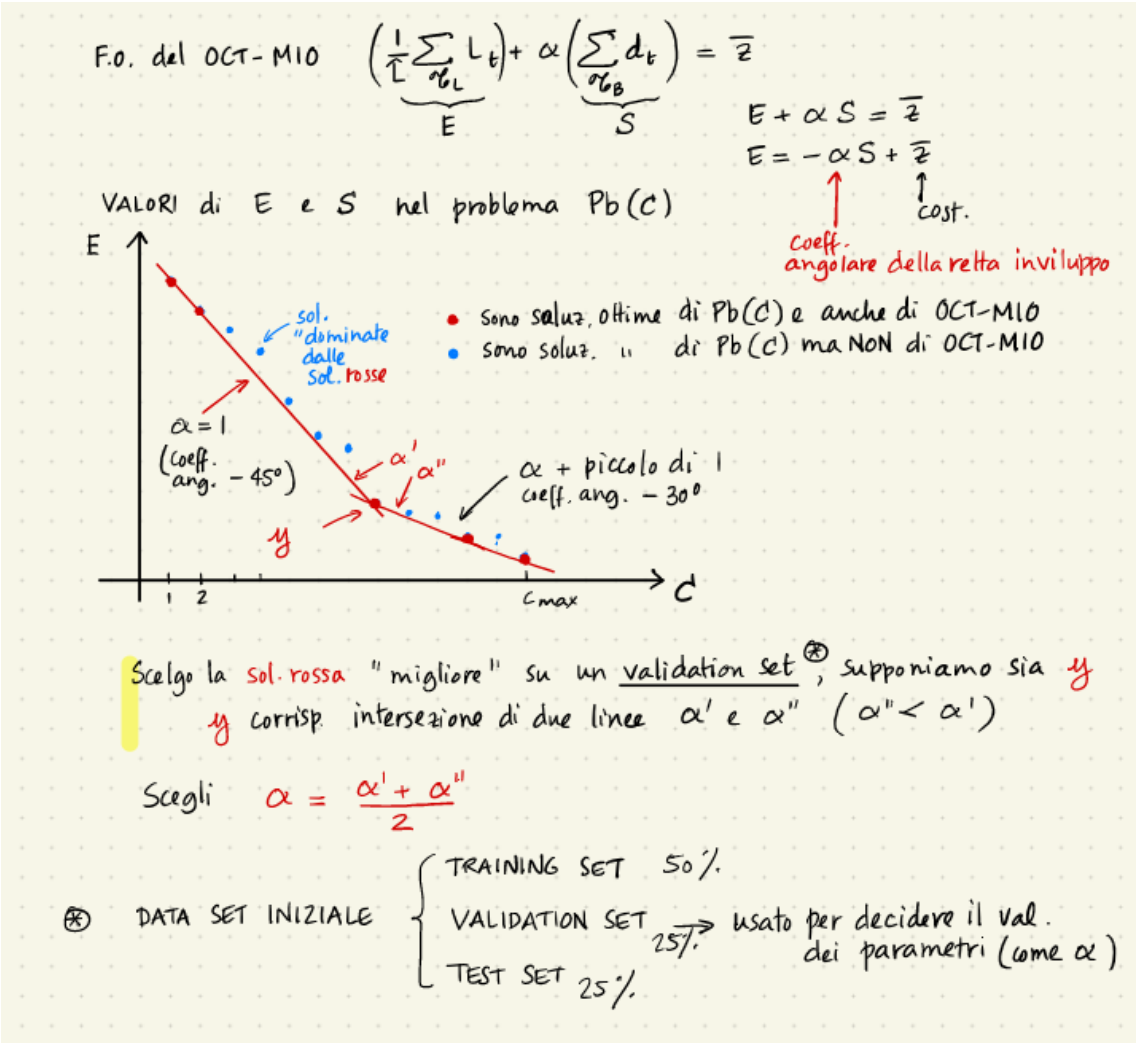
Nell'addestramento di un OCT occorre scegliere la **prodondità** dell'alber $D_{\max}$ da cui si genereranno da $D = 1$(3 nodi) fino a $D_{\max}$.

Possiamo riscrivere il problema di minimizzazione nel seguente modo. Posto $E :=$ errore di training $S :=$ complessità dell'albero:

$$\min E$$

$$S \leqslant C$$

con C è una costante che corrisponde al massimo numero di split $(2^D - 1)$.



Algoritmo 2

1. Scegli $D_{\max}$ e $N_{\min}$;
2. For D=1,..,$D_{\max}$ do:

- For C = $1,\ldots,2^D-1$ do:
  - Run TDITD con $\alpha=0$ e $N_{\min}$. Effettua prune a profondità D,max numero di split = C. Inserisci la soluzione nel pool di warm start:
  - Scegli il candidato più accurato sul validation set nel pool di warm start;
  - Risolvi Pb(C) con profondità D e C split usando il warm start selezionato. Inserisci la soluzione nel pool di warm start;

3. Post-process: rimuovi le soluzioni non ottime per OCT-MIO per alcun valore di α. Determina quindi le soluzioni non dominate e rimuovi le altre.
4. Selezione le soluzioni migliori sul validation set e determina il range di α.

6 OCT: Modelli Multivariati

6.1 Introduzione

Fino ad ora abbiamo considerato alberi di decisione che utilizzano una singola variabile nei loro split ad ogni noto. Questi alberi vengono detti **alberi decisionali univariati.** Ora cerchiamo di estendere questo insieme di alberi decisionali al caso **multivariato**, cioè nel caso in cui negli split di ogni nodo si prendono in considerazione più variabili. Questi alberi decisionali multivariati, vengono detti **OCT-H**.

6.2 Formulazione di OCT-H

Negli alberi decisionali multivariati, non siamo più vincolari a scegliere una singola variabile negli stima, ma possiamo scegliere un generico iperpiano negli split di ogni nodo. Andiamo, ora, a riscrivere i vincoli del nostro problema...

Le variabili a_t sono usate per modellare lo split in ongi nodo ed in particolare $\in[-1,1]^p$:

$$a_{tj} \in [-1,1]$$

$$b_t \in [-1,1]$$

Come nel caso univariato, le variaibli d_t rappresentano la presenza o meno di uno split.

$$\sum_{j=1}^{p} |a_{tj}| \leqslant d_t$$

La problematica di questo vincolo è che è diventato non più lineare. A tale scopo linearizzo tramite delle variabili ausiliarie $\hat{a}_{jt} \geqslant 0 \quad \forall t \in T_B, j=1,\ldots,p$ ottenendo:

$$\hat{a}_{jt} \geqslant a_{jt}; \hat{a}_{jt} \geqslant -a_{jt} \quad j=1,\ldots,p \quad \forall t \in T_B$$

Quindi:

$$\sum_{j=1}^{p} \widehat{a_{jt}} \leqslant d_t$$

Inoltre, abbiamo la variabile b_t che ci indica la presenza di punti dopo la diramazione:

$$-d_t \leqslant b_t \leqslant d_t \quad \forall t \in T_B$$

I **vincoli di consistenza**, mi garantiscono l'assegnamente del punto nel nodo e diventano:

$$\begin{aligned} a_m^T x_i &\geqslant b_m - M(1 - z_{it}) \quad \forall i = 1, \dots, n \quad \forall t \in T_B \quad \forall m \in A_L(t) \\ a_m^T x_i + \mu &\leqslant b_m + M(1 - z_{\text{it}}) \quad \forall i = 1, \dots, n \quad \forall t \in T_B \quad \forall m \in A_R(t) \end{aligned}$$

$\mu = 5 \cdot 10^{-3}$ quantità piccola e di norma si utilizza $M = 2 + \mu$;

$$\begin{aligned} M \leqslant \max\left(a_m^T x_i - b_m\right) &= 2 \\ a_m^T x_i + \mu &\leqslant b_m + (2 + \mu)(1 - z_{it}) \quad \forall i = 1, \dots, n \quad \forall t \in T_L \quad \forall m \in A_L(t) \end{aligned}$$

Varia anche la funzione obiettivo poiché la complessità dell'albero varia: vogliamo tenere in considerazione l'uso di test con più attributi.

Introduco $s_{jt} \in \{0, 1\} \quad \forall t \in T_B, j = 1, \dots, p$:

$$s_{jt} = \begin{cases} 1 & \text{se al nodo } t \in T_B \text{ utilizzo } j \text{ per la definizione dello split } t \\ 0 & \text{altrimenti} \end{cases}$$

$$s_{jt} \geqslant |a_{jt}|$$

Nota 128.

Nel caso in cui $a_{jt} \neq 0 \longrightarrow s_{jt} = 1$

Nel caso in cui $a_{jt} = 0 \longrightarrow s_{jt} = 0$

Linearizzando il vincolo non lineare:

$$-s_{jt} \leqslant a_{jt} \leqslant s_{jt} \quad \forall t \in T_B, j = 1, \dots, p$$

Al fine di ottenere una maggiore efficienza si aggiungono altri due vincoli per rendere s_{jt} compatibile con d_{jt}:

$$s_{jt} \leqslant d_t$$

$$\sum_{j=1}^{p} s_{jt} \geqslant d_t$$

Quindi la funzione obiettivo nel caso OCT-H diventa:

$$\min \frac{1}{\hat{L}} \sum_{t \in T_L} L_t + \alpha \sum_{t \in T_B} \sum_{j=1}^{p} s_{jt}$$

6.3 Warm Start OCT-H

Negli OCT-H come Warm start si utilizza un'euristica greedy appartenente alla TDIDT family. Per determinare il migliore split, invece di utilizzare gli impurity index, si utilizza, sui punti "sopravvissuti" al nodo $t \in T_B$ un OCT-H a profondità $D = 1$.

6.4 Tuning degli Iperparametri

Nella scelta degli iperparametrei si utilizza la stessa procedura di OCT con la differenza che:

$$\text{OCT} \longrightarrow C_{\max} = 2^D - 1$$

$$\text{OCT} - H \longrightarrow C_{\max} = p\,(2^D - 1)$$

7 Alberi di Decisione Pro e Contro

La caratteristica degli alberi di decisione è che sono semplici da spiegare/giustificare ai non esperti. Rappresentano in modo più verosimile il processo decisionale umano rispetto ad altri metodi e modellano in maniera immedieta variabili qualitative. Tuttavia, il livello di accuratezza è inferiori e sono poco robusi poiché piccole variazioni nei dati producono cambiamenti significativi nell'albero risultante,

8 Bagging, Random Forests, Boosting

Gli alberi decisionali DT dsoffrono di **varianza elevata:**

- Estrai in modo casuale due o più dataset da una popolazione;
- Ricava alberi di decisione sulla base dei nuovi dataset;
- alberi di decisione molto dissimili;

Per ridurre la varianza si può ricorrere al **bagging** (Bootstrap Aggregation). Questo è un metodo general-purpose utile a ridurre la variazione aumentanto l'accuratezza della predizione tramite l'aggregazione di un insieme di osservazioni.

Idialemnte si ricava un certo numero B di training set della popolazione osservata e se ne considera la media come predizione del modello aggregato.

8.1 Bagging

Si applica l'idea della sezione precedente agli alberi di classificazione. In particolare si costruisce un albero di classificazione a partire da ciascuno dei training set bootstrap $1, \ldots, B$.

A questo punto, si classifica sulla base del voto di maggioranza cioè dalla classe di predizione più frequente tra i risultati ottenuti dai B alberi.

In aggiunto, con questo tipo di tecnica non si effettua pruning degli alberi generati a partre dai training set bootstrap. La varianza viene ridotta prendendo il risultato a maggioranza sulle B classificazioni.

A questo punto occorre definire una grandezza per valutare l'errore che stiamo commettendo tramite bagging. Questa grandezza di errore viene detts **stima out-of-bag**.

Un DT nel bagging usa circa $\frac{2}{3}$ dei dati originali. Quindi $\frac{1}{3}$ delle osservazioni non vengono utilizzate per generar eun singolo DT e proprio questi punti/osservazioni rappresentano **Out-Of-Bag**.

L'idea è quella di utilizzare i risultati prodotti dai DT che hanno x_i OOB per ottenere una predizione su x_i. Si ripete questa procedura per tutti gli n punti $x_i \quad i = 1, \ldots, n$ e si calcola l'errore di classificazione complessivo.

Ricapitolando il bagging aumenta l'accuratezza della predizione al costo di una minore interpretabilità e utilizza l'indice di Gini come misura dell'importanza di una variabile indipendente j:

Algoritmo 3

Per tutti gli alberi T_b dal training set di bootstrap $b = 1, \ldots, B$:
Calcola la diminuzione d_{b_j} dell'indice di Gini associata allo split sulla variabile j
Output=$\frac{1}{B}\sum_b d_{b_j}$ è la misura dell'**importanza relavita della variabile j**

8.2 Random Forests

I random forests sono un miglioramento del bagging basato sulla scelta di alberi meno correlati tra loro. In particolare, si costruiscono B alberi di decisione e ad ogni split:

- si considerano soltanto un **campione casuale** di $m < p$ variabili candidate;
- La scelta di p viene effettuata in cross validation, ma sperimentalmente si è visto che si pone $p \approxeq \sqrt{m}$

La conseguenza della riduzione delle alternative ad ogni split è che si riducono i rischi dovuti alla forte correlazione tra alberi e si aumenta la riduzione della varianza promuovendo la costruizione di alberi poco correlati tra loro.

In particolae, ponendo $m = p \longrightarrow$ bagging

Con un numero alto di attributi fortemente correlati, scegliere m piccolo rispettoa p può migliorare le performance di queste teniche rispetto al bagging.

8.3 Boosting

Le tecniche d **boosting** possono essere applicabili a metodi diversi di apprendimento statistico di regressione e classificazione; ogni albero viene creato a partire dalle informazoni sugli alberi creati in precedenza e si combina un numero sufficientemente elevato di B alberi di decisione a partire da B traingin set.

In questo caso, i training set non sono bootstrap, ma sono creati modificando in maniera opportuna i dati originali.

I campioni successivi sono determinati pesando i dati dei campioni precedenti.

Il peso di una osservazione x_i al passo b è maggiore se il DT non ha classificato correttamento x_i al passo $(b-1)$. La classifixazione viene effettuata pesando opportunamente i risultati dei vari DT.

8.4 Adaboost ST

Assumo che le label $\in\{-1, 1\}$ ed ho M weak learners k_j a stamp:

$$C(x_i) = \sum_{j=1}^{M} \alpha_j k_j(x_i) \quad x_i \in \text{TS} \quad x_i \in \mathbb{R}^p$$

La risposta:

$$y = \frac{C(x_i)}{|C(x_i)|}$$

Il suo segno determina la risposta del sistema e il denominatore è proporzionale alla confidenza, cioè maggiore è il suo valore, maggiore sarà la confidenza della risposta per il punto del TS.

Inoltre:

$$C_m(x_i) = \sum_{y=1}^{M} \alpha_j k_j(x_i) \longrightarrow C(x_i) = C_M(x_i); C_m(x_i) = C_{m-1}(x_1) + \alpha_m k_m(x_i)$$

A questo punto si necessiata di una funzione per determinare l'errore di C_m:**exponential loss**.

8.5 Errore di classificazione al passo m-esimo

$$E = \sum_{i=1}^{n} e^{y_i C_m(x_i)} \equiv \sum_{i=1}^{m} e^{-y_i C_{m-1}(x_i)} e^{\alpha_i k_m(x_i) y_i} = \sum_{i=1}^{m} W_i^{(m)} e^{\alpha_i k_m(x_i) y_i}$$

8.6 Scelta del Weak Learner al passo m

Il mio obiettivo è quello di avere E minimo:

$$E = \sum_{i:K_m(x_i) \neq y_i}^{m} W_i^{(m)} e^{-\alpha_m k_m(x_i) y_i} + \sum_{i:K_m(x_i) = y_i}^{m} W_i^{(m)} e^{\alpha_m k_m(x_i) y_i} =$$

$$= \sum_{I^=} W_i^{(m)} e^{-\alpha_m} + \sum_{I^{\neq}} W_i^{(m)} (e^{\alpha_m} - e^{-\alpha_m})$$

In cui il primo termine non dipenda da $k_m(\cdot)$, mentre il secondo non dipende ne da i ne da $k_m(\cdot)$. Quindi, per minimizzare viene scelto maniera tale che:

$$\sum_{I^{\neq}} W_i^{(m)}$$

sia minimo.

8.7 Scelta del parametro α_m

Voglio minimizzare E come forma di α_m. Quindi per trovare il minimo derivo $\frac{\partial E}{\partial \alpha_m} = 0$:

$$E = \sum_{i \in I^{\neq}}^{m} W_i^{(m)} e^{-\alpha_m k_m(x_i) y_i} + \sum_{i \in I^=}^{m} W_i^{(m)} e^{-\alpha_m k_m(x_i) y_i} = \sum_{i \in I^{\neq}}^{m} W_i^{(m)} e^{\alpha_m} + \sum_{i \in I^{\neq}}^{m} W_i^{(m)} e^{-\alpha_m} =$$

$$= e^{\alpha_m} W^E - e^{-\alpha_m} W^C$$

Quindi:

$$\frac{\partial E}{\partial \alpha_m} = 0 \Longleftrightarrow e^{\alpha_m} W^E = e^{-\alpha_m} W^C$$

$$\alpha_m + \ln W^E = -\alpha_m + \ln W^C$$

$$\alpha_m = \frac{1}{2} \ln\left(\frac{W^C}{W^E}\right) = \frac{1}{2} \ln\left(\frac{1-\varepsilon}{\varepsilon}\right)$$

Rappresenta un errore complessivo pesato che può essere scritto come:

$$\varepsilon = \frac{W^E}{W^E + W^C}$$

2 Classificazione Robusta

1 Introduzione

Nel caso in cui ci fosse incertezza sui dati, si ha un peggioramento delle prestazioni del classificatore. A tale scopo vogliamo delle procedure che rendano il classificatore più robuto rispetto a delle piccole incertezze. L'approccio standard è la **regolarizzazione** che si basa sulle funzioni di penalità per diminuire l'overfitting. UN altro approccio è quello di utilizzare l'**ottimizzazione robusta.**

2 Robust Classification

Il problema di ottimmizzazione è del timo:

$$\max c(x, y)$$

$$g(x, u) \leqslant 0^m$$

$$x \in X$$

In cui:x

- $x :=$ variabili di decisione;
- $u =$ parametri;
- $g(x, u) \leqslant 0^m :=$ vincoli di diseguaglianza.

I **parametri** u possono essere:

- fissati$\longrightarrow$problema di ottimizzazione incerti;
- incerti$\longrightarrow u \in U$ insieme di incertezaò

.Cerchiamo la soluzione ottima nel caso peggiore.

3 Approccio MaxMin

$$\max_{x \in X} \min_{u \in U} \{c(x, u) : g(x, u) \leqslant 0^m\}$$

Se $|U| = +\infty$ allora si hanno infiniti vincoli che pososno essere ridotti ad un numero finito.

Definizione 129. *Problema Norma Duale*

La norma duale è un numero reale associato ad una funzione lineare definita sullo spazio vettoriale X.

Se $X = \mathbb{R}^n$: una funzione lineare generia $f = a^T x$

$$z_q = \sup \{|f(x)|: \|x\|_q < 1\}$$

$$z_q = \max \{a^T x: \|x\|_q \leqslant 1, x \in \mathbb{R}^n\}$$

Si può dimostrare che la soluzione di questo problema è:

$$z_q = \|a\|_{q^*} \quad \text{con } q^* = \frac{q}{q-1} \quad \text{se } q \neq 1$$

$$z_q = \|a\|_\infty \quad \text{Se } q = 1$$

Si può estenedere questo problema nella norma di $f(x)$ ristretta da un qualsiasi numero $\rho > 0$. Ottenendo:

$$\begin{aligned} z_q &= \max \{a^T x: \|x\|_q \leqslant \rho: x \in \mathbb{R}^n\} = \\ &= \max \{a^T \rho\, y: \|y\|_q \leqslant 1, y \in \mathbb{R}^n\} = \\ &= \rho \|a\|_{q^*} \end{aligned}$$

4 Uncertainty Set

L'**insieme di incertezza** può essere utilizzato per modellare **l'incertezza delle feautre.** Poiché l'incertezza nelle feature è dovuto a mancanza id dati, errori di manipolazioni e errori di misura vogliamo costruire un modello di ottimizzazione per il classificatore che sia una **controparte robusta** del modello OCT.

A tale scopo, supponiamo che:

$$x_i \in \mathbb{R}^p \text{ è il valore esatto}$$

$$(x_i + \Delta x_i) \quad \forall i = 1, \ldots, n \quad \text{Training data effettivo}$$

con $\Delta x_i :=$ perturbazione dell'$i-$esimo dato.

A questo punto, possiamo definire un insieme delle perturbazioni:

$$\Delta X = \{\Delta x_1, \ldots, \Delta x_n\} \in \mathbb{R}^{p \times n}$$

Ottenendo:

$$U_x = \{\Delta x \in \mathbb{R}^{p \times n}: \|\Delta X_i\|_q \leqslant \rho, i = 1, \ldots, n\}$$

Il termine ρ viene detto **parametro di magnitudo(incertezza)** e viene scelto tramite cross validation.

5 Modello MIP (Programmazione Intera Mista)

Se consideriamo un albero binario non compelto, esso sara formato da $\left\lceil \frac{\#D}{2} \right\rceil$. Fissata la struttura, il numero di nodi T è dispari ed abbiamo:

$$\#\text{Foglie} = \left\lceil \frac{T}{2} \right\rceil$$

I nodi sono numerati in modo che $k = 1, 2, \ldots, \left\lfloor \frac{T}{2} \right\rfloor, \left\lceil \frac{T}{2} \right\rceil, \ldots, T$.

Al poso delle vairabili $d_t = \left\{ \begin{array}{l} 1 \text{ split nel nodo } t \\ 0 \text{altirmenti} \end{array} \right.$, si utilizza:

$$\delta_k = \left\{ \begin{array}{l} 1 \text{ se non si ha lo split al nodo } k \\ 0 \text{ altrimenti} \end{array} \right.$$

5.1 Vincoli Strutturali

1. $\delta_k = 1 \quad \text{k}=\left\lceil \frac{T}{2} \right\rceil, \ldots, T;$
2. $\delta_k \geqslant \delta_n \quad k = 1, \ldots, T : u \in A(k);$
3. $\delta_k + \sum_{i=1}^{P} a_{k,i} = 1 \qquad k = 1, 2, \ldots, T$

5.2 Allocazione Punti nelle Foglie

$$z_{i,k} \in \{0, 1\} \quad i = 1, \ldots, n \quad k = 1, \ldots, T$$

Importante 130. $z_{\text{ik}} = 1 \Longleftrightarrow$ assegno punto i al nodo k

Vincoli

- $\sum_{k=1}^{T} z_{ik} = 1 \qquad \text{i=1,..,n}$
- $z_{ik} \leqslant \delta_k \qquad i = 1, \ldots, n \quad k = 1, \ldots, T$
- $\sum_{i=1}^{n} z_{ik} \geqslant N_{\min} l_k \quad l_k \in \{0, 1\} \quad k = 1, \ldots, T \quad l_k = 1$ se la foglia popolata da punti
- $z_{ik} + \delta_u \leqslant 1 \quad i = 1, \ldots, n; k = 1, \ldots, T; u \in A(k)$
- $l_k + \sum_{u \in A(k)} \delta_u \geqslant \delta_k \qquad k = 1, \ldots, T.$

5.3 Funzione Obiettivo

Assumiamo 2 classi $y_i \in \{-1; +1\} \quad \forall i = 1, \ldots, n$

Variabili

$g_k, h_k \in \mathbb{Z}_{\geqslant 0} (\mathbb{R} \geqslant 0)$ contano il numero di punti assegnati $a\, k$ che assumo ± 1

$$(\text{numero di} -1 \text{ in } k) \qquad g_k = \frac{1}{2} \sum_{i=1}^{n} z_{ik}(1 - y_i) \qquad \forall k \in T$$

$$(\text{numero di } 1 \text{ in } k) \quad h_k = \frac{1}{2} \sum_{i=1}^{n} z_{ik}(1 + y_i)$$

Vincoli:

$$(\alpha)\, f_k \leqslant g_k + M\,[w_k + (1 - l_k)] \qquad (\gamma)\, f_k \geqslant -M[1 - w_k + (1 - l_k)]$$

$$(\beta)\, f_k \leqslant h_k + M[1 - w_k + (1 - l_k)] \qquad (\varepsilon)\, f_k \geqslant h_k - M[w_k + (1 - l_k)]$$

Se $l_k = 0 \longrightarrow$ vincoli soddisfatti

Se $l_k = 1$:

$$w_k \in \{0, 1\} \qquad k = 1, \ldots, T$$

Se $w_k = 0 \quad \beta e \gamma$ sono soddisfatti banalmente, ma i vincoli attivi α, ε impongono:

$$h_k \leqslant f_k \leqslant g_k$$

Se $w_k = 1$ $\alpha.\varepsilon$ vengono soddisfatti e rimangono attivi gli altri due e impongono:

$$g_k \leqslant f_k \leqslant h_k$$

Quindi l'insieme f_k contiene tutti i punti malclassificati e quindi può rappresentare la funzione da minimizzare:

$$\min \left[\sum_{k=1}^{T} f_k + \alpha \sum_{k=1}^{T} (1 - \delta_k) \right]$$

Coerenza dei Test

$$a_u^T x_i \leqslant b_u + M\,(1 - z_{ik}) \quad u \in A_l(k) \quad \forall i = 1, \ldots, l$$

$$a_u^T x_i \geqslant b_u - M\,(1 - z_{ik}) \quad u \in A_R(k) \quad \forall i = 1, \ldots, l$$

6 Robustezza vs. Incertezza delle feauture

$$U_x = \{\Delta x \in \mathbb{R}^{p \times n} \colon \|\Delta x_i\|_q \leqslant \rho . i = 1, \ldots, n\}$$

Nei vincoli 5.3 della coerenza dei testi, **controparte robusta**:

$$a_u^T (x_i + \Delta x_i) + \varepsilon \leqslant b_u + M(1 - z_{ik}) \quad \forall i, k, u \in A_L(k) \quad \forall \Delta x \in U_x$$

Possiamo riscriverlo come:

$$a_u^T \Delta x_i \leqslant -a_u^T x_i - \varepsilon + b_u + M(1 - z_{ik})$$

Quindi:

$$\max_{\Delta x \in U_x} \{a_u^T \Delta x_i\} \leqslant -a_u^T - \varepsilon + b_u + M(1 - z_{ik})$$

$$\Updownarrow$$

$$\max_{\|\Delta x = \rho\|} \{a_u^T\, \Delta x \colon \|\Delta x_i\| \leqslant \rho\} = \rho \|a_u\|_q = \rho$$

La versione robustificata diventa:

$$a_u^T x_i + \varepsilon + \rho \leqslant b_u + M(1 - z_{ik})$$

7 Robustezza vs. Incertezza nelle label

7.1 Dualita PL

Il problema primale di minizzazione in un contesto di classificazione è il seguente:

$$z = \min c^T x$$

$$A\, x \leqslant b$$

$$x \in \mathbb{R}^n_{\geqslant 0}$$

Il suo duale:

$$w = \max u^T b$$

$$u^T A \leqslant c^T$$

$$u \in \mathbb{R}^n_{\leqslant 0^m}$$

Se $\overline{x} \in \mathbb{R}^n_{\geqslant 0}$ tale che $A\,\overline{x} \leqslant b$, cioè ammissibile per il primale e $\overline{u} \in \mathbb{R}^m_{\leqslant 0}$ tale che $\overline{u}^T A \leqslant c^T$, ammissibile per il problema duale allora:

$$c^T \overline{x} \geqslant z = w \geqslant \overline{u}^T b$$

7.2 Robustezza nelle label

Se vi è incertezza nelle label, esse sono affette da errore e quindi occorre robustificare il modello e definire l'insieme di incertezza. Sostanzialmente esse sono dei **flip**: la label da 1 diventa -1 e viceversa. Quindi definiamo:

$$\Delta y_i = \begin{cases} 0 & \text{Se non vi è flip} \\ 1 & \text{altrimenti} \end{cases}$$

Posso definire l'insieme di incertezza come:

$$U_y = \left\{ \Delta y_i \in \{0,1\}^n : \sum_{i=1}^{n} \Delta y_i \leqslant \Gamma \right\} \quad \Gamma \in \mathbb{Z}_+$$

La funzione obiettivo, tenendo conto dellìincertezza sulle label diventa:

$$g_k = \frac{1}{2} \sum_{k=1}^{n} z_{ik}(1 - y_i(1 - 2\,\Delta y_i))$$

$$h_k = \frac{1}{2} \sum_{i=1}^{n} z_{ik}(1 + y_i(1 - 2\,\Delta y_i))$$

I vincolI:

$$(\alpha)\, f_k \leqslant g_k + M\,[w_k + (1 - l_k)] \qquad (\gamma)\, f_k \geqslant -M[1 - w_k + (1 - l_k)]$$

$$(\beta) f_k \leqslant h_k + M[1 - w_k + (1 - l_k)] \qquad (\varepsilon)\, f_k \geqslant h_k - M[w_k + (1 - l_k)]$$

Considerando $\alpha + \beta + \gamma + \varepsilon \longrightarrow \text{Se } l_k = \begin{cases} h_k \leqslant f_k \leqslant g_k & \text{se } w_k = 0 \\ g_k \leqslant f_k \leqslant h_k & \text{se } w_k = 1 \end{cases}$

Prendiamo in considerazione solo il primo vincolo, le restanti si svolgono allo stesso modo $\forall k = 1,\ldots, T, \Delta y_i \in U_y$

$$f_k \leqslant \frac{1}{2} \sum_{i=1}^{n} z_{ik}(1 - 2\,\Delta y_i) + M[w_k + (1 - l_k)]$$

$$\frac{1}{2} \sum_{i=1}^{n} (1 - y_i)\, z_{ik} + \sum_{i=1}^{n} z_{ik} y_i\, \Delta y_i \geqslant f_k - M[w_k + (1 - l_k)]$$

$$\frac{1}{2} \sum_{i=1}^{n} (1 - y_i) z_{ik} + \min_{\Delta y \in U} \{z_{ik} y_i\, \Delta y_i\} \geqslant f_k - M[w_k + (1 - l_k)]$$

La funzione di minimo è un problema di programmazione lineare PL, che può essere rilassato:

$$\xi_k = \min_{\Delta y \in U} \{z_{ik} y_i \, \Delta y_i\}$$

$$\sum_{i=1}^{n} \Delta y_i \leqslant \Gamma$$

$$\Delta y_i \in \{0, 1\} \quad i = 1, \ldots, n$$

Che rilassando il vincolo binario:

$$\xi_k = \min_{\Delta y \in U} \{z_{ik} y_i \, \Delta y_i\}$$

$$\sum_{i=1}^{n} \Delta y_i \leqslant \Gamma$$

$$0 \leqslant \Delta y_i \leqslant 1 \quad i = 1, \ldots, n$$

Quindi, risolto questo problema, otteniamo ξ_k:

$$\frac{1}{2} \sum_{i=1}^{n} (1 - y_i) z_{ik} + \xi_k \geqslant f_k - M[w_k + (1 - l_k)]$$

7.3 Proprietà

$$\xi_k^R \leqslant \xi_k \quad \forall k = 1, \ldots, T$$

Allora posso considerare il duale del rilassamento:

$$\begin{aligned} \xi_k = \min_{\Delta y \in U} \{z_{ik} y_i \, \Delta y_i\} \qquad & \tau_k = \max \left(\Gamma \nu_k + \sum_{i=1}^{n} \mu_{ik} \right) \\ \sum_{i=1}^{n} \Delta y_i \leqslant \Gamma \iff & \mu_{ik} + \nu_k \leqslant z_{ik} \, y_i \\ 0 \leqslant \Delta y_i \leqslant 1 \quad i = 1, \ldots, n \qquad & \mu_{ik} \leqslant 0; \nu_k \leqslant 0 \end{aligned}$$

Se sono ammissibili per il duale allora:

$$\left(\Gamma \overline{\nu_k} + \sum_{i=1}^{n} \overline{\mu_{ik}} \right) \leqslant \tau_k \leqslant \xi_k^R \leqslant \xi_k$$

Allora 7.2 diventa:

$$\frac{1}{2} \sum_{i=1}^{n} (1 - y_i) z_{ik} + \left(\Gamma \overline{\nu_k} + \sum_{i=1}^{n} \overline{\mu_{ik}} \right) \geqslant f_k - M[w_k + (1 - l_k)]$$

$$\mu_{ik} + \nu_k \leqslant z_{ik} \, y_i \quad \forall i = 1, \ldots, n$$

$$\mu_{ik} \leqslant 0 \qquad \forall k = 1, \ldots, T$$

$$\nu_k \leqslant 0$$

Avendo così, in conclusione, $n \cdot T$ vincoli invece di T.