

STATISTICA INFERNENZIALE

(Utilizziamo dei dati per inferire (dedurre) delle informazioni sulla popolazione da cui il campione è estratto)

POPOLAZIONE → tutti gli elementi del dataset

Campione → una o più osservazioni della popolazione
↳ campione viene scelto casualmente

↳ Simple Random Sample (SRS)

Parametro → valore caratteristico di una popolazione (μ, σ)

Statistica → valore misurabile dalle caratteristiche di un campione (\bar{x}, s)

Distribuzione dei campioni → La distribuzione di una statistica (misurata sui campioni)

Statistica Inferenziale → stimare uno o più parametri della popolazione utilizzando la statistica dei campioni

POPOLAZIONE	CAMPIONE
Media	\bar{x}
dev. standard	s

SRS → Casuale

→ Rappresentativo → deve rispettare le caratteristiche della POPOLAZIONE

→ Non troppo piccolo

→ NO Bias → non ci devono essere distorsioni rispetto alla statistica da misurare

Distribuzioni campionarie

Media Campionaria

Caso in cui abbiamo un SRS(n) con distribuzione Normale (μ, σ)

↳ media campionaria (\bar{x}) → è una variabile aleatoria con distribuzione normale ($\mu, \sigma/\sqrt{n}$) (distribuzione di \bar{x})

→ Standardizzazione della media campionaria (\bar{x}) $Z \sim N(0,1)$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \text{quando } n \rightarrow \infty$$

Varianza campionaria S^2

In una distribuzione aleatoria normale (μ, δ)

\Rightarrow Varianza campionaria S^2 :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

↙ ↘ ↙

Teorema del limite centrale

Data una SRS(n) con una distribuzione qualunque di popolazione (μ, δ), la distribuzione delle medie dei campioni sarà approssimativamente normale, quando $n \rightarrow \infty$.

Stima dei parametri

La statistica inferenziale cerca di trarre conclusione su un'intera popolazione, basandosi sull'analisi di un sottoinsieme più piccolo (campione).

Il problema è che i veri valori che descrivono una popolazione sono quasi sempre sconosciuti. Questi valori sconosciuti sono chiamati parametri θ . La stima puntuale $\hat{\theta}$ è il metodo che usiamo per indovinare il valore più plausibile di un parametro basandoci su dati del campione che abbiamo raccolto.

Uno stimatore puntuale $\hat{\theta}$ → è una regola o una formula che usiamo per calcolare la stima. È una funzione matematica che viene applicata ai dati del campione.

Per fornire una maggiore affidabilità alla stima puntuale $\hat{\theta}$, aggiorniamo un intervallo di valori possibili, detto intervallo di confidenza, ottenuto a partire da valori che "misurano" il grado di affidabilità della stima.

- ↪ Intervallo piccolo → stima + affidabile
- ↪ Intervallo grande → stima - affidabile

Ottimale se $\theta = \hat{\theta}$ ✓ campione considerato

Spesso → $\hat{\theta}$ → variabile aleatoria → varia da campione a campione

↪ $\hat{\theta} > \theta$ per un altro $\hat{\theta} < \theta$

$$\hat{\theta} = \theta + \underline{\text{errore di stima}}$$

↪ Più piccolo è l'errore migliore è l'estimatore

Errore quadratico medio (MSE) fra $\hat{\theta}$ e θ

$$\text{MSE} = \frac{\sum_{i=1}^n (\hat{\theta} - \theta)^2}{n}$$

Nel confronto fra due stimatori $\hat{\theta}_1$ e $\hat{\theta}_2 \rightarrow$ migliore è quello che ha l'errore quadratico medio minore.

Stimatore puntuale non distorto

↪ Uno stimatore si dice non distorto se il suo valore atteso è esattamente uguale al parametro che stiamo stimando.

$E(\hat{\theta}) \rightarrow$ media della variabile aleatoria $\hat{\theta}$

↪ $E(\hat{\theta}) = \theta$ per ogni possibile valore di θ

HSE

Distorto

↪ Se il valore atteso non è uguale al parametro che stiamo stimando

↪ $E(\hat{\theta}) - E(\theta)$ si dice distorsione di $\hat{\theta}$ (differenza tra il valore atteso dello stimatore e il parametro vero)

Per la proprietà di distorsione dello stimatore \Rightarrow

Principio della stima non distorta

↪ Quando si deve scegliere fra diversi stimatori, scegliere quello non distorto.

↪ Se non ci sono si sceglie in base alla varianza dello stimatore.

Principio della varianza minimale

↪ Fra tutti gli stimatori non distorti di θ scegliere quello con varianza minima.

↪ Risultato \rightarrow MVUE \rightarrow stimatore non distorto di minima varianza.

Stimatore Media $\rightarrow \bar{X} \rightarrow \frac{1}{n} \sum x_i \rightarrow$ Non distorto ($E(\bar{X}) = \mu$)

Stimatore Varianza $\rightarrow S^2 \rightarrow \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow$ Non distorto ($E(S^2) = \sigma^2$)

Stimatore di Varianza Distorto $\rightarrow P^2 \rightarrow \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow$ Distorto ($E(P^2) = \frac{n-1}{n} \sigma^2$)

Intervalli di confidenza (POPOLAZIONE NORMALE)

Stimando con la media campionaria la media della distribuzione, si comette un errore non noto.

$\alpha \rightarrow [0,1] \rightarrow$ livello di confidenza

$$\mu=0 \quad \delta=1$$

$z_{\alpha/2} \rightarrow$ quantile di indice $\alpha/2$ della distribuzione normale standard

↳ definiscono quanto "lontano" dalla stima puntuale possiamo andare per avere una certa possibilità che il parametro vero (es. μ) sia incluso nell'intervallo.

Intervallo di confidenza (IC) (I_α)

↳

$$I_\alpha = \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

↓

intervallo di confidenza $100(1-\alpha)\%$ di μ

Un intervallo di confidenza garantisce un range di valori per un parametro della popolazione, con un certo livello di fiducia.

Considerazioni:

- ↳ Per un livello di confidenza fissato $1-\alpha$, se n aumenta, IC diminuisce
- ↳ Per un n fissato, se $1-\alpha$ aumenta, IC aumenta

Esempio

$$\alpha=0,05 \rightarrow 100(1-\alpha)\% = 95\%$$

↳ Siamo confidenti al 95% che l'intervallo contenga la vera media della popolazione

Se la dev. stand. non è nota.

↳ Dev. std. campionaria $\rightarrow S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

OSSERVAZIONI

$n \rightarrow$ grande

↳ $S \rightarrow$ dev. std. esatta

$n \rightarrow$ piccolo

↳ Sostituisco z con $t \rightarrow$ quantile della distribuzione di t-Student

$$\bar{X} \pm t_{\alpha/2} \left(df = n-1 \right) \frac{S}{\sqrt{n}}$$

Caso popolazione Non Normale

SRS(n) con distribuzione qualunque

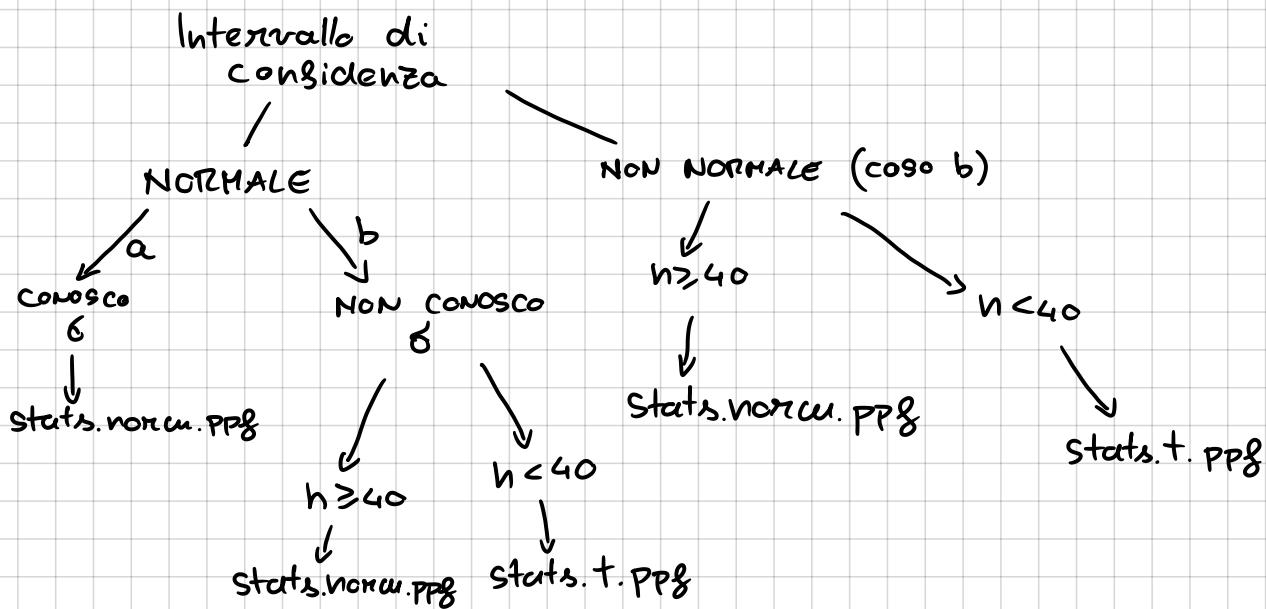
Avere \bar{X} come media campionaria

⚠ Se n sufficientemente grande \bar{X} si comporta come nel caso di campioni estratti da distribuzione normale.

E' difficile determinare la dev. std. esatta di distribuzioni non normali, quindi si utilizza S (dev. std. campionaria)

$$\text{IC} \quad \text{NON NORMALE} \quad \bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Python



Test di Ipotesi

Step del test di ipotesi:

- 1) Considero un SRS(n)
- 2) Formulo un H_0 (ipotesi nulla)
- 3) Formulo una H_1 (ipotesi alternativa)
- 4) Calcolo una statistica
- 5) Confronto il valore della statistica calcolata con quello dell'ipotesi nulla e calcolo un valore detto p-value
- 6) Interpretro il p-value e decido se l'ipotesi nulla e' da rifiutare oppure no

Esempio

la cacciazione dei voti degli studenti considerato

Hipotesi nulla H_0 $\mu_0 = 105$

Hipotesi alternativa $H_a \rightarrow$ contraddice l'ipotesi nulla

$H_0: \mu = \mu_0$ ipotesi nulla

$H_a: \mu < \mu_0$ ipotesi alternativa

Risultato $\rightarrow H_0$ rigettata (in favore di H_a)
 $\rightarrow H_0$ non rigettata

Se H_0 non è rigettata non significa che non c'è sufficientemente evidenza che H_a sia vera, cioè $\mu < \mu_0$

P-Value

Risultato
 test di ipotesi $\rightarrow p < 0,05 \Rightarrow H_0$ rigettata
 $\rightarrow p \geq 0,05 \Rightarrow H_0$ non rigettata

Tipi di errore

1) Errori del primo tipo (Producer risk errors)

$\hookrightarrow H_0$ è vera \rightarrow ma la statistica dice no (H_0 sarebbe vera)

2) Errori del secondo tipo (consumer risk errors)

$\hookrightarrow H_0$ accettata \rightarrow la statistica conferma

\hookrightarrow problema: si accettano misure non corrette

Sensitività, Specificità, PPV e NPV

Sensitività \rightarrow positivi correttamente identificati da un test

Specificità \rightarrow negativi correttamente identificati da un test

Positive Predictive Value (PPV) \rightarrow test positivi che sono stati correttamente predetti

Negative Predicted Value (NPV) \rightarrow test negativi che sono stati correttamente predetti

		Condition		Test Outcome	Positive predictive value = $\frac{\Sigma \text{True Positive}}{\Sigma \text{Test Outcome Positive}}$		
		Condition Positive					
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)				
	Test Outcome Negative	False Negative (Type II error)	True Negative				
		$\text{Sensitivity} = \frac{\Sigma \text{True Positive}}{\Sigma \text{Condition Positive}}$	$\text{Specificity} = \frac{\Sigma \text{True Negative}}{\Sigma \text{Condition Negative}}$		$\text{Negative predictive value} = \frac{\Sigma \text{True Negative}}{\Sigma \text{Test Outcome Negative}}$		
		Sensibilità	Specificità				

Prevalenza e Incidenza

Risponde alla domanda, se il test risulta positivo, qual'è la probabilità che sia vero?

Test di ipotesi: prevalenza e incidenza

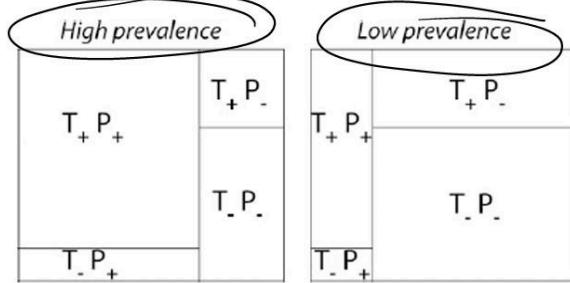


Fig. 7.8 Effect of prevalence on PPV and NPV. "T" stands for "test," and "P" for "patient." (For comparison with below: $T+P+ = TP$, $T-P- = TN$, $T+P- = FP$, and $T-P+ = FN$.)

Esempio test di Hp:

False positive rate (α)

↳ Errore di tipo I: Nel nostro test.

Rifiuta l'ipotesi nulla (H_0) quando in realtà H_0 è vera.

$$\text{Type Error 1} = 1 - \text{Specificity} = \frac{FP}{FP+TN}$$

		Condition		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive (TP) = 25	False Positive (FP) = 175	Positive predictive value = $\frac{TP}{TP+FP} = 25 / (25+175) = 12.5\%$
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 2000	Negative predictive value = $\frac{TN}{FN+TN} = 2000 / (10+2000) = 99.5\%$
	Sensitivity = $\frac{TP}{TP+FN} = 25 / (25+10) = 71\%$	Specificity = $\frac{TN}{TN+FP} = 2000 / (175+2000) = 92\%$		

- È la probabilità che il test sia positivo dato che realemente positivo.

- Quanto pesano i falsi positivi su tutti gli effettivi negativi ($FP+TN$)

False negative rate (β)

↳ Errore di tipo II: Nel nostro test. Non rifiuta l'ipotesi nulla (H_0) quando in realtà è falsa (H_1 vera)

$$\text{Type Error 2} = 1 - \text{Sensitivity} = \frac{FN}{TP+FN}$$

- È la probabilità che il test sia negativo dato che realemente positivo.
- Quanto pesano i falsi negativi su tutti gli effettivi positivi.

$$\text{Positive likelihood ratio} = \frac{\text{sensitivity}}{1 - \text{specificity}} = 8.9$$

- Quante volte è più probabile ottenere un test positivo quando la condizione è vera rispetto a quando è falsa.

$$\text{Negative likelihood ratio} = \frac{1 - \text{sensitivity}}{\text{specificity}} = 0.3$$

- Quante volte è più probabile ottenere un test negativo quando la condizione è vera rispetto a quando è falsa.

Test di Normalità

QQ plot → Il quantile del dataset considerato è plottato rispetto al quantile di una distribuzione (noi prendiamo normale).

→ Due distribuzioni sono simili → i punti devono stare vicino alla retta.

Test di Shapiro-Wilk

↪ p-value → se il p-value è:

- $p\text{-value} > 0.05$ Non rifiutiamo H_0
 - $p\text{-value} \leq 0.05$ Rifiutiamo H_0 .
-

Maximum Likelihood Estimation

Il metodo di massima verosimiglianza (MLE) è una tecnica della statistica inferenziale per stimare i parametri di un modello statistico a partire dai dati osservati. L'idea centrale è trovare i valori dei parametri che rendono più probabile (o verosimile) osservare i dati effettivamente raccolti.

È costituito da 2 fasi:

- 1) Costruire la funzione di verosimiglianza L .
- 2) Massimizzare la funzione di verosimiglianza.

Funzione di verosimiglianza

• Formalmente definita come prodotto delle funzioni di densità valutata in ciascun punto del campione.

• Supponiamo che X_1, \dots, X_n sia un SRS(n) → da una distribuzione PMF o PDF

↪ $f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_m)$ → f dipendente da m parametri $\theta_1, \dots, \theta_m$.
 ↓

La sua funzione di Verosimiglianza è:

$$L(\theta_1, \dots, \theta_m) = \prod_{i=1}^n f(x_i; \theta_1, \dots, \theta_m)$$

L → è una funzione dei parametri θ . I dati x_i sono fissati e noti.

Perché la permutatoria (\prod)? → Le osservazioni sono indipendenti. La probabilità congiunta di osservare tutto il campione sotto ipotesi di indipendenza è la \prod .

Come la risolviamo? → Fissiamo i valori osservati del campione (x_1, \dots, x_n) e consideriamo la funzione f non più come funzione della variabile x , ma come funzione dei parametri θ .

Funzione di Verosimiglianza e MLE:

Distribuzione binomiale

$p \rightarrow$ probabilità

$$L(p) = p^{\sum_i x_i} (1-p)^{n - \sum_i x_i}$$

Si deve stimare il parametro di probabilità p di una distribuzione binomiale.

Distribuzione normale

$\mu \rightarrow$ media

$\sigma \rightarrow$ deviazione standard

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\sum_i (x_i - \mu)^2 / \sigma^2}$$

Vogliamo stimare i parametri μ e σ con uno stimatore MLE.

- Successivamente per stimare i parametri di una distribuzione bisogna massimizzare la funzione di verosimiglianza.

Massimizzazione di una funzione di Verosimiglianza

• Non è detto che esista il massimo di L e non è detto che sia unico

• Spesso conviene minimizzare $- \ln L(\theta)$

→ anziché massimizzare

$$\rightarrow \operatorname{argmin}_x g(x) = \operatorname{argmin}_x -f(x)$$

→ Spesso si ricorre ad algoritmi ottimizzatori

Funzioni a Più Variabili

Sia $f(x, y)$ una funzione definita in un insieme aperto $A \subset \mathbb{R}^2$ e sia $P_0 = (x_0, y_0)$ un punto di A ...

Possiamo definire due rapporti incrementali:

Rapporto incrementale rispetto a x
$$\frac{f(x_0, y_0) - f(x_0, y_0)}{x - x_0}$$

Rapporto incrementale rispetto a y
$$\frac{f(x_0, y) - f(x_0, y_0)}{y - y_0}$$

Se esiste ed è finito il limite per $x \rightarrow x_0$ del primo rapporto, si dice la funzione $f(x, y)$ è parzialmente derivabile rispetto a x nel punto $P_0 = (x_0, y_0)$

Rapporto incrementale parziale: isolano l'effetto di una variabile mantenendo l'altra fissa.

Derivata parziale
rispetto a x nel punto
 $P_0 = (x_0, y_0)$

$$\frac{\delta f}{\delta x}(x_0, y_0) \quad f_x(x_0, y_0)$$

Analogamente se esiste il limite per $y \rightarrow y_0$

$g_x(x_0, y_0) \rightarrow$ e' la pendenza della retta tangente alla curva ottenuta inserendo la superficie $z = g(x, y)$ con il piano $x = x_0$ nel punto P_0

- La derivata parziale \rightarrow ci fornisce la variazione istantanea di quota della funzione $g(x, y)$ rispettivamente rispetto alla variabile x e alla y .

Derivate parziali: seconde

Se $g_{xx}(x, y)$ e' derivabile, e' possibile calcolarne le derivate parziali rispetto ad x e ad y

$$g_{xx}(x, y) \rightarrow \frac{\partial^2 g}{\partial x^2}(x, y) \rightarrow \frac{\partial}{\partial x} \frac{\partial g}{\partial x}$$

$$g_{xy}(x, y) \rightarrow \frac{\partial^2 g}{\partial y \partial x}(x, y) \rightarrow \frac{\partial}{\partial y} \frac{\partial g}{\partial x}$$

$$g_{yx}(x, y) \rightarrow \frac{\partial^2 g}{\partial x \partial y}(x, y) \rightarrow \frac{\partial}{\partial x} \frac{\partial g}{\partial y}$$

$$g_{yy}(x, y) \rightarrow \frac{\partial^2 g}{\partial y^2}(x, y) \rightarrow \frac{\partial}{\partial y} \frac{\partial g}{\partial y}$$

Ci possono essere 4 derivate seconde parziali:

Gradiente di una funzione

Se esistono in (x_0, y_0) la derivata parziale rispetto ad x e la derivata parziale rispetto ad y , che indichiamo con $g_x(x_0, y_0)$ e $g_y(x_0, y_0)$.

E' possibile costruire un vettore che ha per componenti le derivate parziali:

Gradiente della funzione g valutato in (x_0, y_0)

$$\nabla g(x_0, y_0) = (g_x(x_0, y_0), g_y(x_0, y_0))$$

Naba

Funzioni a 3 variabili

$$\nabla g(x_0, y_0, z_0) = (g_x(x_0, y_0, z_0), g_y(x_0, y_0, z_0), g_z(x_0, y_0, z_0))$$

$$\text{es. } \nabla g(1, 0) = (2, 0)$$

Nel punto $(1, 0)$ il modo piu' rapido per salire sulla superficie e' spostarsi lungo la direzione dell'asse x positivo. Perche' la componente x del gradiente e' positiva mentre y e' 0.

Massima Decrescenza \rightarrow Il vettore opposto $-\nabla g = (-2, 0)$ punta lungo l'asse x negativo. Significa che la discesa piu' rapida dal punto $(1, 0)$ e' spostarsi verso l'origine $(0, 0)$.

Ottimizzazione

Gli algoritmi numerici per calcolare il massimo di una funzione in più variabili vengono detti algoritmi di ottimizzazione.

$$\min_x f(x)$$

$x \in \mathbb{R}^n$ è un vettore reale di $n \geq 1$ componenti e la funzione obiettivo $f: \mathbb{R}^n \rightarrow \mathbb{R}$ è una funzione regolare.

La funzione obiettivo: serve a misurare le prestazioni di un sistema o la qualità di una soluzione. Il suo scopo è fornire un criterio numerico, un punteggio per confrontare diverse alternative e dire quale sia "migliore" o "peggiore".

Così:

x^* → vettore

Minimo Locale $f(x^*) \leq f(x) \quad \forall x \text{ in un intorno di } x^*$

Minimo in senso stretto $f(x^*) < f(x) \quad \forall x \text{ in un intorno di } x^*$

Direzione in discesa $(\nabla f(x^*), d) < 0$

Minimo globale $f(x^*) \leq f(x) \quad \forall x \in \mathbb{R}^n$

Massimo globale $f(x^*) \geq f(x) \quad \forall x \in \mathbb{R}^n$

Massimo Locale $f(x^*) \geq f(x) \quad \forall x \text{ in un intorno di } x^*$

Teorema (Condizioni necessarie del primo ordine)

→ \exists tutte le derivate prime parziali di f

Se x^* è un punto di minimo locale e f è differenziabile con continuità in un intorno aperto di x^* , allora $\nabla f(x^*) = 0$.

$\nabla f(x^*) = 0 \rightarrow$ punto stazionario

x^* punto di minimo $\Rightarrow \nabla f(x^*) = 0$



Condizione Necessaria Non Sufficiente

La Condizione $\nabla f(x^*) = 0$ è condizione necessaria anche x^* sia un punto di minimo locale, tale condizione non è però sufficiente poiché si tratta di un punto stazionario qualsiasi.

Algoritmi di ottimizzazione

↳ Algoritmi iterativi $\rightarrow x_{k+1} = G(x_k)$

↳ Per la minimizzazione di una funzione in generale NON convergono al minimo globale, ma ad un minimo locale.

Metodi di discesa

Il vettore p è una direzione di discesa di f in x se esiste un $\bar{\alpha} > 0$ t.c.:

$$f(x + \alpha p) \leq f(x) \quad \forall \alpha \in]0, \bar{\alpha}]$$

- La direzione dell'antigradiente $p = -\nabla f(x)$ è sempre in direzione di discesa.

Criterio d'arresto

x_0 input $x_0, x_1, \dots, x_k, x_{k+1} \rightarrow x^*$

$K=0$ $\left[\begin{array}{l} \text{finché criterio vero} \\ x_{k+1} = G \\ K = K+1 \end{array} \right]$

Differenza tra Algoritmi di ottimizzazione Vincolata e Non Vincolata

A. Ottimizzazione Vincolata \rightarrow Le variabili decisionali devono soddisfare una serie di restrizioni (vincoli). Questi vincoli definiscono una regione accettabile, un sottoinsieme di R^n all'interno del quale dobbiamo cercare la soluzione ottima.

A. Ottimizzazione Non Vincolata \rightarrow Nell'ottimizzazione non vincolata non ci sono restrizioni sui valori che le variabili decisionali possono assumere. L'obiettivo è trovare il punto di minimo o massimo della funzione obiettivo esplorando tutto lo spazio R^n .

\hookrightarrow In questo caso per criterio di corretto si intende il criterio che dovrebbe indicare il raggiungimento con successo di un punto stazionario con tolleranza specificata dall'utente.

- Negli Algoritmi di Ottimizzazione esistono diversi criteri per garantire le iterazioni e che possono indicare il raggiungimento di una soluzione oppure il fallimento dell'algoritmo