Università degli Studi di Palermo

Facoltà di Economia

CdS Statistica per l'Analisi dei Dati

Appunti del corso di Matematica

27 - Funzioni di più Variabili – Ottimizzazione

Anno Accademico 2013/2014

M. Tumminello e A. Consiglio

1. Introduzione

Nella parte relativa alle serie di funzioni, abbiamo visto che possiamo sviluppare in serie di Taylor una funzione, se la funzione ammette derivate di tutti gli ordini in un punto c. Lo sviluppo in serie non è altro che un modo per approssimare una funzione tramite polinomi di grado $n-$esimo. In particolare, si ipotizzi uno sviluppo in c fino al termine di secondo grado:

$$f(c+h) = f(c) + f'(c)\,h + \frac{1}{2}f''(c)h^2 + o(h^2, c),$$

dove $o(h^2, c)$ è una funzione di h e c tale che:

$$\lim_{h\to 0} \frac{o(h^2, c)}{h^2} = 0.$$

Quindi $o(h^2, c)$ è una quantità che tende a 0 più velocemente di h^2 al tendere di h a 0.

Come possiamo generalizzare questa approssimazione al secondo ordine nel caso di funzioni a più variabili? In effetti, abbiamo già visto l'approssimazione al primo ordine quando abbiamo introdotto il concetto di derivabilità di una funzione di più variabili in un punto:

$$f(\overrightarrow{c} + \overrightarrow{h}) \simeq f(\overrightarrow{c}) + \overrightarrow{\nabla} f(\overrightarrow{c}) \cdot \overrightarrow{h}$$

(nella notazione usata in precedenza era $\overrightarrow{x} = \overrightarrow{c} + \overrightarrow{h}$, da cui $\overrightarrow{h} = \overrightarrow{x} - \overrightarrow{c}$).
Il seguente teorema fornisce l'approssimazione di Taylor al secondo ordine di una funzione di più variabili.

Teorema 1.1. *Sia $f : \mathbb{R}^n \to \mathbb{R}$ una funzione di classe C^2 in un intorno di $\overrightarrow{c}$, $B_r(\overrightarrow{c})$. Sia, inoltre, $\overrightarrow{x} = \overrightarrow{c} + \overrightarrow{h} \in B_r(\overrightarrow{c})$. Allora esiste un numero reale $s \in [0, 1]$ tale che:*

$$f(\overrightarrow{c} + \overrightarrow{h}) = f(\overrightarrow{c}) + \overrightarrow{\nabla} f(\overrightarrow{c}) \cdot \overrightarrow{h} + \frac{1}{2}\overrightarrow{h}^T H[f(\overrightarrow{c} + s\,\overrightarrow{h})]\,\overrightarrow{h},$$

dove $H[f(\overrightarrow{c} + s\,\overrightarrow{h})]$ è la matrice Hessiana di f calcolata nel punto $\overrightarrow{c} + s\,\overrightarrow{h}$).

Se si pone $\overrightarrow{x} = \overrightarrow{c} + \overrightarrow{h}$ e si calcola l'Hessiano in $\overrightarrow{c}$ il membro destro dell'uguaglianza precedente diventa l'approssimazione polinomiale al secondo ordine di f in un intorno di $\overrightarrow{c}$.

1.1. Definizione di Polinomio di Taylor al Secondo Ordine.

Sia $f : \mathbb{R}^n \to \mathbb{R}$ una funzione di tipo C^2 in un intorno di $\overrightarrow{c}$, $B_r(\overrightarrow{c})$.

Si definisce **Polinomio di Taylor del Secondo Ordine** di $f(\cdot)$ in $\overrightarrow{c}$ la seguente espresssione:

$$P_2(\overrightarrow{x}) = f(\overrightarrow{c}) + \overrightarrow{\nabla} f(\overrightarrow{c}) \cdot (\overrightarrow{x} - \overrightarrow{c}) + \frac{1}{2}(\overrightarrow{x} - \overrightarrow{c})^T\, H[f(\overrightarrow{c})]\,(\overrightarrow{x} - \overrightarrow{c}).$$

Esempio 1.1
Determinare il polinomio di Taylor del secondo ordine della funzione

$$f(x,y) = e^{-2\,x+y}$$

in $\overrightarrow{c} = (0,0)$.

Si ha:

$$\overrightarrow{\nabla} f(\overrightarrow{x}) = \left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y)\right) = \left(-2\,e^{-2\,x+y}, e^{-2\,x+y}\right);$$

$$H[f(\overrightarrow{x})] = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x,y) & \frac{\partial^2 f}{\partial x\,\partial y}(x,y) \\ \frac{\partial^2 f}{\partial y\,\partial x}(x,y) & \frac{\partial^2 f}{\partial y^2}(x,y) \end{bmatrix} = \begin{bmatrix} 4\,e^{-2\,x+y} & -2\,e^{-2\,x+y} \\ -2\,e^{-2\,x+y} & e^{-2\,x+y} \end{bmatrix}.$$

Nel punto $\overrightarrow{c} = (0,0)$ otteniamo: $f(0,0) = 1$, $\overrightarrow{\nabla} f(0,0) = (-2,1)$ e $H[f(0,0)] = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$.

quindi:

$$\begin{aligned} P_2(x,y) =& 1 + \begin{bmatrix} -2 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2}\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \\ =& 1 - 2\,x + y + 2\,x^2 - 2\,x\,y + \frac{1}{2}\,y^2. \end{aligned}$$

In Fig.1, a pagina seguente, è riportato il grafico della funzione $f(x,y)$ e della sua approssimazione polinomiale a secondo ordine, $P_2(x,y)$, in un intorno di $\overrightarrow{c} = (0,0)$.

2. Ottimizzazione

L'ottimizzazione di funzioni a più variabili gioca un ruolo fondamentale nella metodologia statistica. Dai minimi quadrati alla massima verosimiglianza, la determinazione di massimi e minimi coincide con la ricerca dei migliori stimatori.

Introduciamo alcune definizioni che descrivono in maniera formale cosa è un ottimo di una funzione di più variabili.

2.1. Definizione di Massimo e Minimo.

Sia $f : K \subseteq \mathbb{R}^n \to \mathbb{R}$ una funzione.

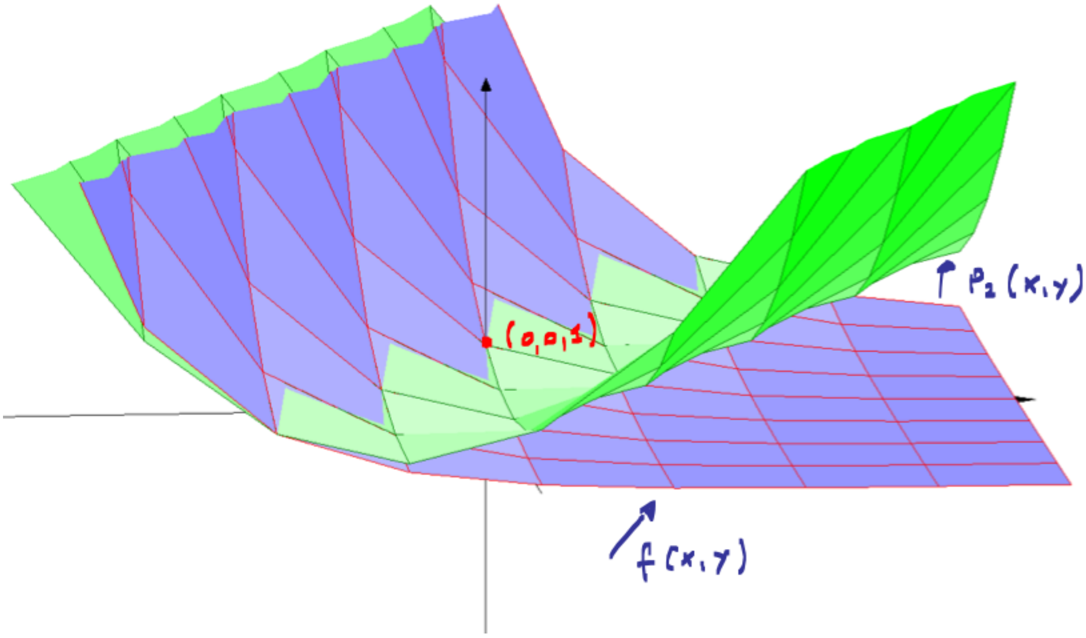

FIGURE 1. Rappresentazione 3D della funzione $f(x,y) = e^{-2\,x+y}$ e della sua approssimazione polinomiale $P_2(x,y) = 1 - 2\,x + y + 2\,x^2 - 2\,x\,y + \frac{1}{2}\,y^2$ in un intorno di $(0,0)$.

(1) Un punto $\overrightarrow{M} \in K$ è un **massimo** di $f(\overrightarrow{x})$ su K se $\forall \overrightarrow{x} \in K$ allora $f(\overrightarrow{M}) \geq f(\overrightarrow{x})$.
(2) Un punto $\overrightarrow{M} \in K$ è un **massimo locale (o massimo relativo)** di $f(\overrightarrow{x})$ se esiste un intorno $B_r(\overrightarrow{M})$ tale che $\forall \overrightarrow{x} \in K \cap B_r(M)$ allora $f(\overrightarrow{M}) \geq f(\overrightarrow{x})$.

Chiaramente, le definizioni di **minimo** e **di minimo locale** si ottengono invertendo il senso delle disuguaglianze.

(1) Un punto $\overrightarrow{m} \in K$ è un **minimo** di $f(\overrightarrow{x})$ su K se $\forall \overrightarrow{x} \in K$ allora $f(\overrightarrow{m}) \leq f(\overrightarrow{x})$.
(2) Un punto $\overrightarrow{m} \in K$ è un **minimo locale (o minimo relativo)** di $f(\overrightarrow{x})$ se esiste un intorno $B_r(\overrightarrow{m})$ tale che $\forall \overrightarrow{x} \in K \cap B_r(m)$ allora $f(\overrightarrow{m}) \leq f(\overrightarrow{x})$.

Di solito, per enfatizzare il fatto che un punto di massimo (minimo) è tale rispetto a tutto il dominio, secondo la definizione (1), ci si riferisce ad $\overrightarrow{M}$ come ad un **punto di massimo globale** o **punto di massimo assoluto** e a $\overrightarrow{m}$ come ad un **punto di minimo globale** o **punto di minimo assoluto**.

Per le funzioni $f : \mathbb{R} \rightarrow \mathbb{R}$ abbiamo visto che le condizioni di primo ordine, $f'(x) = 0$ opurre $f'(x)$ che non esiste, permettono di identificare i cosiddetti punti critici o stazionari. Vediamo come si generalizzano queste condizioni per funzioni $f : \mathbb{R}^n \to \mathbb{R}$.

TEOREMA 2.1 (Condizioni Necessarie o di Primo Ordine). *Sia $f : K \subseteq \mathbb{R}^n \to \mathbb{R}$ una funzione di classe C^1 su K. Condizione necessaria (NON sufficiente) affinché un punto $\overrightarrow{x}_0 \in K$ sia di massimo oppure di minimo relativo per f è che*

$$\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$$

PROOF. Consideriamo un punto $\overrightarrow{x}_0 \in K$ e un suo intorno sferico $B_r(\overrightarrow{x}_0)$. Vogliamo mostrare che la condizione $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$ è necessaria affinché $\overrightarrow{x}_0$ sia un punto di massimo o di minimo per f. Sia $\overrightarrow{x} \in B_r(\overrightarrow{x}_0)$. Consideriamo l'approssimazione lineare di f nell'intorno di $\overrightarrow{x}_0$. Se poniamo $\overrightarrow{x} - \overrightarrow{x}_0 = \overrightarrow{h}$, abbiamo già visto che:

$$f(\overrightarrow{x}) = f(\overrightarrow{x}_0 + \overrightarrow{h}) = f(\overrightarrow{x}_0) + \overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} + o(\overrightarrow{h}).$$

Se $\overrightarrow{x}_0$ è un punto di massimo locale allora $\forall \overrightarrow{x} \in B_r(\overrightarrow{x}_0)$ deve essere $f(\overrightarrow{x}) \leq f(\overrightarrow{x}_0)$ e, quindi, $f(\overrightarrow{x}) - f(\overrightarrow{x}_0) \leq 0$. Dall'equazione precedente abbiamo che:

$$f(\overrightarrow{x}) - f(\overrightarrow{x}_0) = \overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} + o(\overrightarrow{h}).$$

Il segno del membro destro dell'equazione dipende solo dal segno di $\overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h}$ se esso è diverso da zero, poiché il termine restante è tale che $\lim_{\overrightarrow{h} \to \overrightarrow{0}} \frac{o(\overrightarrow{h})}{||\overrightarrow{h}||} = 0$. Se $\overrightarrow{\nabla} f(\overrightarrow{x}_0) \neq 0$ allora $\overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} \neq 0$. Supponiamo che $\overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} > 0$. Allora $f(\overrightarrow{x}) - f(\overrightarrow{x}_0) > 0$. Se consideriamo ora un certo $\overrightarrow{x}' \in B_r(\overrightarrow{x}_0)$ definito come $\overrightarrow{x}' - \overrightarrow{x}_0 = -\overrightarrow{h}$, ovvero consideriamo il vettore opposto a $\overrightarrow{x}$ rispetto a $\overrightarrow{x}_0$ risulterà $f(\overrightarrow{x}') - f(\overrightarrow{x}_0) = -\overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} < 0$. Dunque $\overrightarrow{x}_0$ non può essere un punto di massimo. Con ragionamento analogo si dimostra che $\overrightarrow{x}_0$ non può essere di minimo. Dunque se $\overrightarrow{x}_0$ è di massimo, oppure di minimo, deve essere $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$. □

Si osservi che, in generale, le condizioni del primo ordine corrispondono alla soluzione di un sistema quadrato di equazioni non lineari. La condizione mostrata nel precedente teorema è solo necessaria affinché un punto sia di massimo o di minimo. Infatti, se $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$ allora $f(\overrightarrow{x}) - f(\overrightarrow{x}_0) = \overrightarrow{\nabla} f(\overrightarrow{x}_0) \cdot \overrightarrow{h} + o(\overrightarrow{h}) = 0 + o(\overrightarrow{h})$ e, quindi, il segno di $f(\overrightarrow{x}) - f(\overrightarrow{x}_0)$ dipenderà da quello di $o(\overrightarrow{h})$. Questo ci porta a stabilire le cosiddette condizioni sufficienti o al secondo ordine. Premettiamo prima una definizione.

2.2. Definizione di punti critici o stazionari. Sia $f : K \subseteq \mathbb{R}^n \to \mathbb{R}$ una funzione di classe C^1 su K e $\overrightarrow{x}_0 \in K$. Si dice che $\overrightarrow{x}_0$ è **punto critico** o **punto stazionario** di f se $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$.

Teorema 2.2 (Condizioni sufficienti o di secondo ordine). *Sia $f : K \subseteq \mathbb{R}^n \to \mathbb{R}$ una funzione di classe C^2 su K. Sia inoltre $\overrightarrow{x}_0$ un punto critico di f, ovvero $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$. Il punto $\overrightarrow{x}_0$ è un punto di minimo (massimo) locale di f se la matrice Hessiana $H[f(\overrightarrow{x}_0)]$ è definita positiva (negativa). Il punto $\overrightarrow{x}_0$ è un punto di sella se $H[f(\overrightarrow{x}_0)]$ non è definita. Nel caso, infine, in cui tutti gli autovalori di $H[f(\overrightarrow{x}_0)]$ sono nulli, oppure $H[f(\overrightarrow{x}_0)]$ è semidifinita positiva o negativa, non si può dire nulla.*

Proof. Dimostriamo per un punto di minimo. Consideriamo un punto critico $\overrightarrow{x}_0$ di f. Sia inoltre $\overrightarrow{x} = (\overrightarrow{x}_0 + \overrightarrow{h}) \in B_r(\overrightarrow{x}_0)$. Il teorema di Taylor permette di approssimare f nell'intorno di $\overrightarrow{x}_0$ al secondo ordine:

$$f(\overrightarrow{x}) = f(\overrightarrow{x}_0+\overrightarrow{h}) = f(\overrightarrow{x}_0)+\overrightarrow{\nabla} f(\overrightarrow{x}_0)\cdot \overrightarrow{h} + \frac{1}{2}\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h} + r(\overrightarrow{h})$$

dove il termine $r(\overrightarrow{h})$ è il resto dello sviluppo di Taylor a secondo ordine ed è, in valore assoluto sempre minore del secondo termine dello sviluppo se questo è diverso da zero: $|r(\overrightarrow{h})| < \frac{1}{2}|\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h}|$. Se $\overrightarrow{x}_0$ è un punto critico, allora $\overrightarrow{\nabla} f(\overrightarrow{x}_0) = \overrightarrow{0}$. Quindi:

$$f(\overrightarrow{x}_0 + \overrightarrow{h}) = f(\overrightarrow{x}_0) + \frac{1}{2}\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h} + r(\overrightarrow{h}),$$

ovvero,

$$f(\overrightarrow{x}_0 + \overrightarrow{h}) - f(\overrightarrow{x}_0) = \frac{1}{2}\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h} + r(\overrightarrow{h}),$$

Il punto $\overrightarrow{x}_0$ è un minimo se $\forall \overrightarrow{x} \in B_r(\overrightarrow{x}_0)$ allora $f(\overrightarrow{x}) - f(\overrightarrow{x}_0) \geq 0$. Abbiamo:

$$f(\overrightarrow{x}) - f(\overrightarrow{x}_0) = f(\overrightarrow{x}_0 + \overrightarrow{h}) - f(\overrightarrow{x}_0) = \frac{1}{2}\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h} + r(\overrightarrow{h}),$$

Poichè abbiamo detto che $|r(\overrightarrow{h})| < \frac{1}{2}|\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h}|$ allora il segno dell'ultimo membro sarà dato dal segno di $\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h}$. Dunque $f(\overrightarrow{x}) - f(\overrightarrow{x}_0) \geq 0$ se $\overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h} \geq 0$. Questa condizione richiede che la forma quadratica $Q(\overrightarrow{h}) = \overrightarrow{h}^T H[f(\overrightarrow{x}_0)]\,\overrightarrow{h}$ sia definita positiva e dunque che la matrice Hessiana $H[f(\overrightarrow{x}_0)]$ sia definita positiva. □

Esempio 2.1
Sia $Q(\overrightarrow{x}) = \overrightarrow{x}^T A \overrightarrow{x}$ una forma quadratica con:

$$A = \begin{bmatrix} 5 & -1 & -2 \\ -1 & 5 & 4 \\ -2 & 4 & 5 \end{bmatrix}.$$

Determinare i punti stazionari di $Q(\overrightarrow{x})$ e stabilire se sono punti di minimo, massimo o sella.

Abbiamo già visto come calcolare il gradiente di una forma quadratica: $\overrightarrow{\nabla} Q(\overrightarrow{x}) = 2A\overrightarrow{x}$. I punti critici sono punti in cui $\overrightarrow{\nabla} Q(\overrightarrow{x}) = \overrightarrow{0}$. Dobbiamo dunque risolvere il sistema lineare omogeneo $2A\overrightarrow{x} = \overrightarrow{0}$ che è equivalente a risolvere $A\overrightarrow{x} = \overrightarrow{0}$:

$$\begin{bmatrix} 5 & -1 & -2 \\ -1 & 5 & 4 \\ -2 & 4 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Lo studente dimostri che questo sistema ammette come unica soluzione quella banale: $\overrightarrow{x} = (x_1, x_2, x_3) = (0, 0, 0)$. Questo è dunque l'unico punto critico della forma quadratica $Q(\overrightarrow{x})$. Dobbiamo ora calcolare la matrice Hessiana di $Q(\overrightarrow{x})$. Tenuto conto del fatto che, per una forma quadratica, $\overrightarrow{\nabla} Q(\overrightarrow{x}) = 2A\overrightarrow{x}$, ovvero

$$\overrightarrow{\nabla} Q(\overrightarrow{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = 2A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

lo studente mostri che $H[Q(\overrightarrow{x})] = 2A$. Si noti che, nel caso di forme quadratiche, la matrice Hessiana è costante (non dipende dal punto in cui le derivate di secondo ordine sono calcolate). Quindi per capire se $\overrightarrow{x}_0 = (0, 0, 0)$ è un punto di massimo, minimo o sella dobbiamo determinare la definitezza di A. Lo studente mostri che la matrice A è definita positiva (ad esempio mostrando che tutti i suoi autovalori sono positivi). Dunque il punto $\overrightarrow{x}_0 = (0, 0, 0)$ è un punto di minimo. Cosa sarebbe successo se $|A| = 0$? Si consideri l'esempio seguente.

Esempio 2.2
Sia $Q(\overrightarrow{x}) = \overrightarrow{x}^T A \overrightarrow{x}$ una forma quadratica la cui matrice associata è:

$$A = \begin{bmatrix} 4/3 & 2 \\ 2 & 3 \end{bmatrix}.$$

Determinare i punti stazionari di $q(\overrightarrow{x})$ e stabilire se sono punti di minimo, massimo o sella.

Ragionando come nell'esempio precedente si ottiene che il sistema $\overrightarrow{\nabla} Q(\overrightarrow{x}) = \overrightarrow{0}$ ammette infinite soluzioni, in quanto la matrice dei coefficienti ha determinante nullo. In particolare le soluzioni del sistema sono rappresentate dal sottospazio di $\mathbb{R}^2$ generato dal vettore $\overrightarrow{u} = (-3/2, 1)$, quindi una retta del piano. Questa retta rappresenta tutti gli infiniti punti critici della forma quadratica. Lo studente dimostri, analizzando la matrice Hessiana, che questi punti critici sono punti di minimo, come mostra la seguente Fig.2.

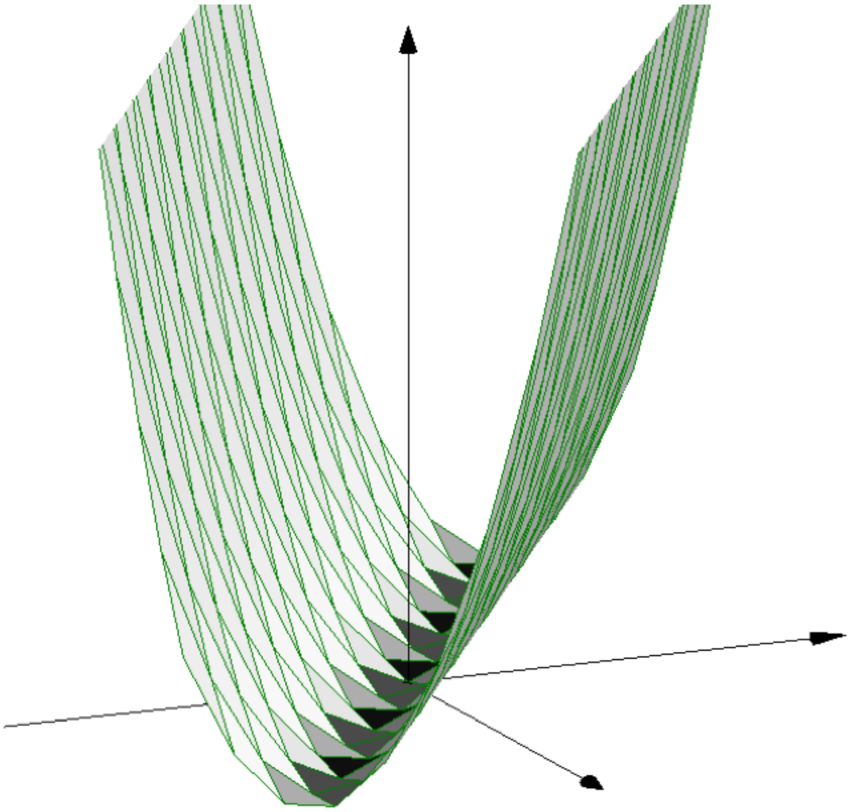

FIGURE 2. Rappresentazione 3D della forma quadratica considerata nell'Esempio 2.2.

Vediamo ora un'applicazione statistica del metodo di ottimizzazione introdotto. Si ipotizzi che esista una relazione che lega una variabile z, detta variabile **dipendente**, ad altre due variabili, x e y, dette variabili **indipendenti**. Per esempio, è possibile ipotizzare che esista una relazione tra il peso (variabile dipendente), l'altezza e l'età (variabili indipendenti), per stabilire se un individuo di data altezza ed età sia in sovrappeso. E' evidente che, se questa relazione esiste, non sarà perfetta, non solo a causa di fluttuazioni statistiche, ma anche perché, in molti casi, si preferisce descrivere la relazione tramite funzioni lineari, piuttosto che non lineari. Dato un campione di n misurazioni z_i, y_i e x_i, $\forall i = 1, \cdots, n$, dove n è la dimensione campionaria, si ipotizzi che:

$$z_i = f(x_i, y_i) + \varepsilon_i,$$

dove ε_i è l'errore di minimizzazione o approssimazione, o include entrambi. La forma funzionale f che lega z a x e y può essere di diverso tipo. Per esempio, supponiamo sia del tipo: $z = f(x, y) = \beta_0 + \beta_1 x + \beta_2 y^2$, quindi, una funzione quadratica, oppure, più semplicemente, una funzione lineare $z = f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$. Si noti che entrambe queste funzioni sono comunque lineari rispetto ai parametri, che è ciò che rende semplice il processo di ottimizzazione. Ad esempio, nella figura riportata di seguito (Fig.3), si può dedurre che la relazione tra volume, altezza e circonferenza di un albero sia lineare, ovvero descritta da un piano.

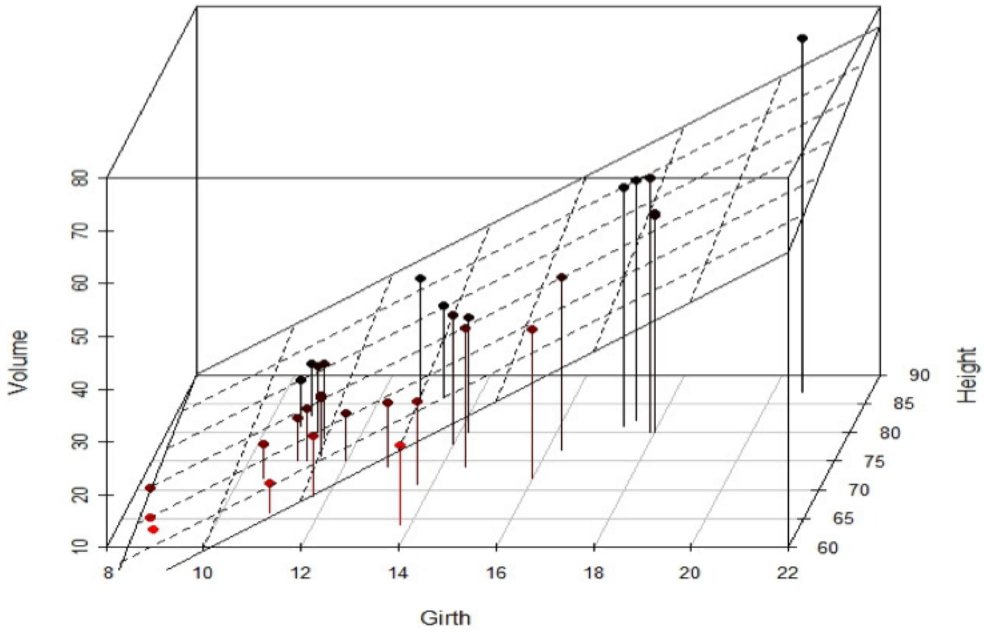

Figure 3. Grafico che riporta i valori di volume, altezza e circonferenza di un set di alberi e riporta un piano a rappresentare la relazione tra queste quantità.

Gli errori ε_i sono rappresentati dalla differenza tra il valore del volume z_i ed il valore corrispondente della $f(x_i, y_i)$. Chiaramente dati i valori di z_i, y_i e x_i, $\forall i = 1, \cdots, n$, i singoli valori di ε_i dipendono dalla scelta dei parametri β_0, β_1 e β_2. Quindi nella fase di stima, possiamo pensare che:

$$\varepsilon_i = z_i - f(\beta_0, \beta_1, \beta_2, x_i, y_i).$$

Come possiamo determinare il "migliore" valore di β_0, β_1 e β_2? Innanzitutto dobbiamo definire cosa intendiamo per "migliore". Una possibile definizione consiste nel definire "migliori" quei valori di β_0, β_1 e β_2 che rendono minima la somma dei quadrati degli errori. Quindi sceglieremo quei valori di β_0, β_1 e β_2 che rendono minima la funzione (dei soli parametri):

$$\begin{aligned} F(\beta_0, \beta_1, \beta_2) &= \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [z_i - f(\beta_0, \beta_1, \beta_2, x_i, y_i)]^2 \\ &= \sum_{i=1}^{n} [z_i - (\beta_0 + \beta_1\, x_i + \beta_2\, y_i)]^2 \end{aligned}$$

Le condizioni necessarie per trovare un punto di minimo richiedono che il gradiente di $F(\beta_0, \beta_1, \beta_2)$ si annulli nel punto. Abbiamo quindi:

$$\vec{\nabla} F(\beta_0, \beta_1, \beta_2) = \begin{bmatrix} \frac{\partial F}{\partial \beta_0} \\ \frac{\partial F}{\partial \beta_1} \\ \frac{\partial F}{\partial \beta_2} \end{bmatrix} = -2 \begin{bmatrix} \sum_{i=1}^{n} [z_i - (\beta_0 + \beta_1\, x_i + \beta_2\, y_i)] \\ \sum_{i=1}^{n} [z_i - (\beta_0 + \beta_1\, x_i + \beta_2\, y_i)] \cdot x_i \\ \sum_{i=1}^{n} [z_i - (\beta_0 + \beta_1\, x_i + \beta_2\, y_i)] \cdot y_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

La ricerca dei punti critici consiste dunque, in questo caso, nella risoluzione del seguente sistema di equazioni lineari

$$\begin{cases} n\,\beta_0 + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i; \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i\, z_i; \\ \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i\, y_i + \beta_2 \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i\, z_i, \end{cases}$$

rispetto a β_0, β_1 e β_2. Tale sistema prende il nome di sistema di equazioni normali. Questo sistema è un sistema di Kramer. Si lascia allo studente ottenere la soluzione del sistema. Si noti che la soluzione è unica. Inoltre, si verifica facilmente che la matrice Hessiana del sistema non dipende da β_0, β_1 e β_2. In particolare:

$$H[F(\beta_0, \beta_1, \beta_2)] = \begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i\, y_i & \sum_{i=1}^{n} y_i^2 \end{bmatrix}$$

E' possibile mostrare che $H[F(\beta_0, \beta_1, \beta_2)]$ è semidefinita positiva e quindi il punto cercato è effettivamente un punto di minimo.

Un metodo alternativo di risolvere il problema di trovare il "migliore" valore di β_0, β_1 e β_2 consiste nel rappresentare le quantità coinvolte come matrici e vettori. Nel caso specifico sia A la matrice dei dati per le variabili indipendenti cui aggiungiamo come prima colonna il vettore di dimensione n (pari alla lunghezza del campione) formato da tutti 1:

$$A^T = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ y_1 & y_2 & y_3 & \cdots & y_n \end{bmatrix}.$$

Siano inoltre $\overrightarrow{\beta} = (\beta_1, \beta_2, \beta_3)$ e $\overrightarrow{z}$ il vettore dei dati relativo alla variabile dipendente: $\overrightarrow{z} = (z_1, z_2, z_3, \cdots, z_n)$. Con questa notazione il vettore degli errori $\overrightarrow{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ può essere scritto come:

$$\overrightarrow{\varepsilon} = \overrightarrow{z} - A \cdot \overrightarrow{\beta}$$

e la somma dei quadrati degli errori non è altro che $||\overrightarrow{\varepsilon}||^2 = \overrightarrow{\varepsilon} \cdot \overrightarrow{\varepsilon} = (\overrightarrow{z} - A \cdot \overrightarrow{\beta}) \cdot (\overrightarrow{z} - A \cdot \overrightarrow{\beta})$. Quindi il problema di minimo risulta trovare il valore di $\overrightarrow{\beta}$ tale che sia minimo il prodotto scalare $(\overrightarrow{z} - A \cdot \overrightarrow{\beta}) \cdot (\overrightarrow{z} - A \cdot \overrightarrow{\beta})$. Per le proprietà del prodotto scalare abbiamo che:

$$\begin{aligned} F(\overrightarrow{\beta}) = (\overrightarrow{z} - A \cdot \overrightarrow{\beta}) \cdot (\overrightarrow{z} - A \cdot \overrightarrow{\beta}) = (\overrightarrow{z} - A \cdot \overrightarrow{\beta})^T (\overrightarrow{z} - A \cdot \overrightarrow{\beta}) = \\ = \overrightarrow{z}^T \overrightarrow{z} - (A\overrightarrow{\beta})^T \overrightarrow{z} - \overrightarrow{z}^T (A\overrightarrow{\beta}) + (A\overrightarrow{\beta})^T (A\overrightarrow{\beta}) = \\ = \overrightarrow{z}^T \overrightarrow{z} - 2\, \overrightarrow{\beta}^T A^T \overrightarrow{z} + \overrightarrow{\beta}^T A^T A\, \overrightarrow{\beta}. \end{aligned}$$

Le condizioni necessarie sono date da:

$$\overrightarrow{0} = \frac{\partial F(\overrightarrow{\beta})}{\partial \overrightarrow{\beta}} = -2\, A^T\, \overrightarrow{z} + 2A^T\, A\, \overrightarrow{\beta}$$

Dunque se la matrice $A^T\, A$ è invertibile possiamo risolvere l'equazione precedente rispetto a $\overrightarrow{\beta}$:

$$A^T\, \overrightarrow{z} = A^T\, A\, \overrightarrow{\beta} \quad \Leftrightarrow \quad \overrightarrow{\beta} = (A^T\, A)^{-1}\, A^T \overrightarrow{z}.$$

Lo studente verifichi che $A^T\, A\, \overrightarrow{\beta} = A^T\, \overrightarrow{z}$ è il sistema di equazioni normali e che $\frac{\partial^2 F(\overrightarrow{\beta})}{\partial \overrightarrow{\beta}^2} = A^T\, A$ coincide con la matrice dei coefficienti del sistema di equazioni normali. Si noti infine che una qualsiasi matrice che possa essere scritta come il prodotto di una matrice per la sua trasposta $(A^T\, A)$ è sempre semidefinita positiva.

Esempio 2.3
Determinare i punti critici della funzione

$$f(x,y) = -\frac{1}{4}\,x^4 + \frac{2}{3}\,x^3 + 4\,x\,y - y^2$$

e verificare quali sono i punti di massimo, di minimo o di sella.

Le condizioni necessarie sono:

$$0 = \frac{\partial f}{\partial x} = -x^3 + 2\,x^2 + 4\,y; \quad 0 = \frac{\partial f}{\partial y} = 4\,x - 2\,y$$

Queste due equazioni devono essere soddisfatte contemporaneamente nei punti critici. Dalla seconda equazione otteniamo $y = 2\,x$ e, sostituendo nella prima equazione, otteniamo:

$$-x^3 + 2\,x^2 + 8\,x = 0 \quad \Leftrightarrow \quad x\,(x-4)\,(x+2) = 2$$

Pertanto abbiamo 3 punti critici, uno per ciascuna soluzione dell'equazione precedente: $(0,0)$, $(4,8)$, $(-2,-4)$. La matrice Hessiana è data da:

$$H[f(x,y)] = \begin{bmatrix} -3\,x^2 + 4\,x & 4 \\ 4 & -2 \end{bmatrix}$$

Quindi, nei tre punti critici, abbiamo:

$$H[f(0,0)] = \begin{bmatrix} 0 & 4 \\ 4 & -2 \end{bmatrix} \quad \textbf{Indefinita (SELLA);}$$

$$H[f(4,8)] = \begin{bmatrix} -32 & 4 \\ 4 & -2 \end{bmatrix} \quad \textbf{Definita negativa (MASSIMO);}$$

$$H[f(-2,-4)] = \begin{bmatrix} -20 & 4 \\ 4 & -2 \end{bmatrix} \quad \textbf{Definita negativa (MASSIMO).}$$

Il grafico della funzione $f(x,y)$ riportato Fig.4 conferma quanto trovato analiticamente.

Esercizi 2.1
Si determinino i punti critici delle funzioni:

$$f(x,y) = x\,\sin(y) \quad \text{e} \quad f(x,y) = x\,y\,\log(x\,y),$$

e si determini se essi sono punti di massimo, di minimo oppure di sella.

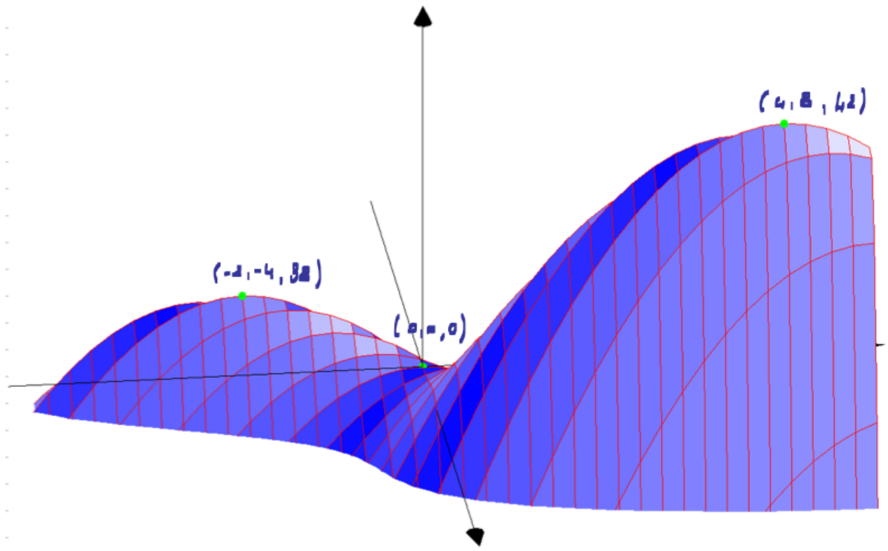

Figure 4. Grafico della funzione $f(x,y) = -\frac{1}{4}\,x^4 + \frac{2}{3}\,x^3 + 4\,x\,y - y^2$.