Localizzazione con reti di sensori wireless mediante regressione Gaussiana

14 novembre 2008

Indice

Prefazione

La *regressione* rappresenta il problema di estrapolare un modello continuo da un set di dati che sia un sottoinsieme dei dati sul quale vogliamo applicare il modello; si tratta perciò di un procedimento di tipo induttivo, in quanto si cerca di ricostruire un modello generale basandosi su un insieme limitato di dati sperimentali. La regressione si inserisce assieme alla classificazione (che ricerca invece un modello discreto) nell'ambito dell'*apprendimento supervisionato.*

Applicheremo questo metodo a dei dati raccolti nel laboratorio *NavLab*, dove l'insieme dei dati campione sarà rappresentato dalle potenze registrate da 11 sensori wireless (ingressi) e dalla posizione del dispositivo mobile (uscita). Lo scopo è creare un modello da cui ricavare la posizione (x, y) del dispositivo mobile, date come ingresso le potenze misurate dai sensori $(x_1, ..., x_{11})$; dove inoltre l'ingresso è particolarmente rumoroso per la natura dell'ambiente e dei sensori.

Il vantaggio di questo tipo di approccio è che il modello viene costruito senza la necessità di una conoscenza approfondita dei fenomeni fisici che governano il dispositivo; per questo motivo è indicato nel caso che il modello abbia una intrinseca difficoltà di formalizzazione, che può essere dovuta sia al numero delle variabili in gioco (ad esempio nelle previsioni meteorologiche), sia alla mancanza di leggi matematiche certe (come nel riconoscimento dei caratteri, oppure nelle previsioni dei mercati finanziari), o ad altre variabilità difficilmente codificabili (ad esempio una sorta di previsto indice di gradimento di una pubblicità on-line, data una serie di siti visitati precedentemente, modello soggettivo che varia da utente a utente). Inoltre, utilizzando i processi gaussiani, si hanno notevoli vantaggi dal punto di vista computazionale.

Capitolo 1

Regressione

1.1 Introduzione alla regressione

Supponiamo di avere una funzione incognita $f(x)$, di campionarla a intervalli casuali e di voler ricostruire questa funzione da questo insieme finito di campioni. Il problema ammette infinite soluzioni plausibili, in quanto sono ammissibili tutte le funzioni che passano per i punti campionati. Inoltre, in un modello più realistico, il campionamento avviene in presenza del rumore e dunque si allarga ulteriormente l'insieme delle soluzioni, che ammette come soluzione anche le funzioni che passano ad una certa distanza (dipendente dal rumore) dai punti campionati. Nel costruire il modello faremo delle assunzioni sulla funzione da ricostruire, ad esempio supporremo sempre che la funzione abbia un certo grado di regolarità, cioè che ad ingressi simili corrispondano uscite simili (correlazione tra gli ingressi).

1.2 Inferenza Bayesiana

Il procedimento dell'inferenza Bayesiana permette di combinare alcune conoscenze del modello che abbiamo "a priori" (cioè le informazioni sul modello che abbiamo prima di vedere i dati sperimentali) con le informazioni che ricaviamo dai campioni.

Il ruolo fondamentale è ricoperto dal Teorema (o regola) di Bayes:

$$p(a|b) = \frac{p(b|a) \cdot p(a)}{p(b)} \tag{1.1}$$

I due tipi di informazione sono rappresentati da distribuzioni di probabilità: la distribuzione *a priori* [1] e la distribuzione *verosimiglianza* [2]. La distribuzione a priori rappresenta dunque le nostre conoscenze iniziali del modello, che, combinate secondo la regola di Bayes con le informazioni contenute nei dati campione (training set) e rappresentate dalla verosimiglianza danno luogo alla distribuzione di probabilità *a posteriori* [3]

1.3 Modello lineare

Prima di mostrare questo procedimento nella costruzione di un modello lineare, introduciamo una notazione. Denotiamo con $\mathcal{D}$ l'insieme degli n campioni $(\mathbf{x}_i, y_i)$; $\mathbf{x}_i$ è un vettore colonna di dimensione D rappresentante l'ingresso e y_i l'uscita. L'insieme X è formato dall'aggregazione degli n vettori $\mathbf{x}_i$, dunque è rappresentabile con una matrice $D \times n$, mentre gli output possono essere rappresentati dal vettore y.

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, i = 1, ..., n\}$$

$$\textit{Rappresentazione matriciale}: \ \mathcal{D} = (X, \mathbf{y})$$

Supponiamo dunque di dover modellare una funzione lineare $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$, i cui valori campionati sono corrotti dal rumore additivo, Gaussiano a variabili indipendenti identicamente distribuite $\epsilon = \mathcal{N}(0, \sigma_n^2)$:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}, \quad y = f(\mathbf{x}) + \epsilon \tag{1.2}$$

Innanzi tutto, dobbiamo specificare la distribuzione a priori, che indica le nostre informazioni sul modello, ad esempio una distribuzione Gaussiana a media zero e matrice covarianza Σ_p:

$$\mathbf{w} = \mathcal{N}(\mathbf{0}, \Sigma_p) \tag{1.3}$$

Siccome abbiamo posto che il rumore sia a variabili indipendenti possiamo fattorizzare la distribuzione di probabilità della verosimiglianza:

[1] Nella letteratura inglese: *prior*
[2] *likelyhood*
[3] *posterior*

$$p(\mathbf{y}|X,\mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i,\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \cdot e^{-\frac{(y_i-\mathbf{x}_i^T\mathbf{w})^2}{2\sigma_n^2}} = \frac{1}{(2\pi\sigma_n^2)^{\frac{n}{2}}} \cdot e^{-\frac{|\mathbf{y}-X^T\mathbf{w}|^2}{2\sigma_n^2}}$$

$$p(\mathbf{y}|X,\mathbf{w}) = \mathcal{N}(X^T\mathbf{w},\, \sigma_n^2) \tag{1.4}$$

A questo punto, utilizzando la regola di Bayes, possiamo calcolare la distribuzione a posteriori:

$$posteriori = \frac{priori \times verosimiglianza}{normalizzazione}, \qquad p(\mathbf{w}|\mathbf{y},X) = \frac{p(\mathbf{w}) \cdot p(\mathbf{y}|X,\mathbf{w})}{p(\mathbf{y}|X)} \tag{1.5}$$

La costante di normalizzazione $p(\mathbf{y}|X)$[4] è indipendente dai parametri $\mathbf{w}$ ed è ottenuta (secondo il teorema della Probabilità Totale) integrando su tutti i possibili parametri:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|X,\mathbf{w}) \cdot p(\mathbf{w}) \tag{1.6}$$

Tralasciando la costante di normalizzazione, otteniamo

$$p(\mathbf{w}|\mathbf{y},X) \propto e^{-\frac{1}{2}(\mathbf{w}^T\Sigma^{-1}\mathbf{w})} \cdot e^{-\frac{1}{2\sigma_n^2}(\mathbf{y}-X^T\mathbf{w})^T(\mathbf{y}-X^T\mathbf{w})} \tag{1.7}$$

dopo alcuni calcoli otteniamo:

$$p(\mathbf{y}|X) \backsim \mathcal{N}(\frac{1}{\sigma_n}A^{-1}X\mathbf{y}A^{-1}), \quad con\ A = \sigma_n^{-2}XX^T + \Sigma_p^{-1} \tag{1.8}$$

Per effettuare una predizione su un punto $\mathbf{x}_*$ calcoliamo la media pesata rispetto a tutti i valori che i parametri possono assumere:

$$\begin{aligned} p(f_*|\mathbf{x}_*,X,\mathbf{y}) &= \int p(f_*|\mathbf{x}_*,\mathbf{w}) \cdot p(\mathbf{w}|X,\mathbf{y}) \\ &= \mathcal{N}(\frac{1}{\sigma_n}\mathbf{x}_*A^{-1}X\mathbf{y},\ \mathbf{x}_*^TA^{-1}\mathbf{x}_*) \end{aligned} \tag{1.9}$$

Siamo cioè andati a vedere la probabilità che il punto $\mathbf{x}_*$ abbia il valore f_* data la scelta di parametri w ($p(f_*|\mathbf{x}_*,\mathbf{w})$) e applicando il teorema della

[4]Nella letteratura inglese: *marginal likelyhood*

probabilità totale abbiamo calcolato la probabilità che il punto $\mathbf{x}_*$ abbia il valore f_* indipendentemente dalla scelta dei parametri. La distribuzione di probabilità, in quanto risultato di prodotti e integrazioni di distribuzioni di probabilità Gaussiane risulta anch'essa Gaussiana; il valore predetto sarà quindi la media[5] in quanto valore più probabile, mentre la varianza ci darà un'indicazione dell'incertezza in quel punto della nostra previsione.

Questo metodo ovviamente darà buoni risultato quando la funzione $f(\mathbf{x})$ è lineare o la si può ritenere tale con buona approssimazione.

1.4 Ampliamento del modello lineare con basi di funzioni

Un modo per ovviare alla mancanza di flessibilità del modello lineare e renderlo adatto a funzioni non lineari è proiettare il modello in uno spazio di funzioni. Usiamo una mappa $\phi(\mathbf{x}) : \mathbb{R}^D \to V$ per proiettare lo spazio degli ingressi di dimensione D in uno spazio di funzioni V di dimensione N, generato da una base di funzioni $< \phi_1, ..., \phi_N >$; se $\phi(\mathbf{x})$ è formata da funzioni non dipendenti da $\mathbf{w}$ il modello è lineare nello spazio V e possiamo quindi ricorrere al procedimento esposto nella sezione precedente.

Il vettore degli ingressi è $\phi(\mathbf{x})$, la matrice X corrisponde ora alla matrice $\Phi(\mathbf{x})$, le cui colonne sono formate dai vettori $\phi(\mathbf{x})$. Il modello da costruire è relativo alla funzione $f(\mathbf{x}) = \phi(\mathbf{x}^T)\mathbf{w}$. Con queste sostituzioni, la distribuzione dei valori predetti (1.9) diventa:

$$p(f_*|x_*, X, y) \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\phi^T(\mathbf{x}_*)A^{-1}\Phi(X)\mathbf{y},\ \phi^T(\mathbf{x}_*)A^{-1}\phi(\mathbf{x}_*)\right) \tag{1.10}$$

Per effettuare una predizione occorrerebbe invertire la matrice A che è di dimensione $(N \times N)$, dal momento che potrebbe essere grande a piacere (dipendendo dalla dimensione dello spazio V scelto) è conveniente riscrivere l'equazione in questo modo: definiamo $K = \Phi^T\Sigma_p\Phi$.

$$f_*|x_*, X, y \sim \mathcal{N}\left(\phi_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\mathbf{y},\ \phi_*^T\Sigma_p\phi_* - \phi_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma_p\phi_*\right) \tag{1.11}$$

In questo modo è necessario invece invertire una matrice di dimensione $(n \times n)$, la cui dimensione dipende dal numero di punti campione del training set.

[5] o mediana o moda, che coincidono in quanto la distribuzione è Gaussiana

1.5 Processi Gaussiani

1.5.1 Metodo

Nel precedente paragrafo, proiettando lo spazio degli ingressi X nello spazio V e considerando $f(\mathbf{x}) = \phi(\mathbf{x}^T)\mathbf{w}$ abbiamo considerato la funzione $f(x)$ come una combinazione lineare di funzioni $\phi_1(\mathbf{x}), ..., \phi_N(\mathbf{x})$, combinate secondo il vettore degli scalari $\mathbf{w}$ che rappresenta il "peso" che attribuiamo a ciascuna funzione, peso che sarà maggiore a seconda della probabilità di realizzazione che attribuiamo a quella funzione. Il procedimento richiede necessariamente una base finita, e quindi di operare in uno spazio V di dimensione finita.

Utilizzando i processi gaussiani si può ottenere un modello equivalente al modello lineare proiettato in uno spazio V di dimensione infinita (cioè generato da una base infinita di funzioni), questo perché il processo gaussiano definisce una distribuzione di probabilità nello spazio delle funzioni. Così come una variabile aleatoria Gaussiana è completamente specificata dalla sua media e varianza, un processo aleatorio gaussiano è specificato dalla funzione media e funzione covarianza:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{1.12}$$

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(x)] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})) \cdot (f(\mathbf{x}') - m(\mathbf{x}'))] \end{aligned} \tag{1.13}$$

Specificando una media $m(\mathbf{x})$ e una funzione covarianza $k(\mathbf{x}, \mathbf{x}')$ stiamo specificando una distribuzione di probabilità a priori nello spazio delle funzioni, stiamo cioè specificando le informazioni che abbiamo sul modello prima di vedere dati campione.

$$\mathbf{f}_* \sim \mathcal{N}(\mathbf{0}, k(X_*, X_*)) \tag{1.14}$$

Poniamo per il momento $m(\mathbf{x}) = 0$, vedremo poi come procedere nel caso di una generica $m(\mathbf{x})$ *deterministica*. Una volta in possesso degli n dati campione ($\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ con $y_i = f(\mathbf{x}_i) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$) dobbiamo "unire" le informazioni che abbiamo a priori sul modello con le informazioni che possiamo ricavare dai dati. La covarianza tra due punti y_p, y_q è data data dalla somma $k(x_p, x_q) + \sigma_n^2 \delta_{pq}$ poiché il rumore ϵ è una variabile aleatori indipendente. Scriviamo quindi la distribuzione di probabilità congiunta:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{1.15}$$

Dalla distribuzione di probabilità congiunta ricaviamo la distribuzione di probabilità a posteriori nei punti di test X_* cioè la probabilità che la funzione valutata nei punti d'ingresso $X_* = \mathbf{x}_{*1}, ..., \mathbf{x}_{*m}$ abbia valori d'uscita $\mathbf{f}_* = y_{*1}, ..., y_{*m}$ dati i punti campione $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$:

$$p(\mathbf{f}_*|X_*, X, \mathbf{f}) \sim \mathcal{N}(\,\overline{\mathbf{f}}_*,\ cov(\overline{\mathbf{f}}_*)\,) \tag{1.16}$$

$$\overline{\mathbf{f}}_* \triangleq \mathbb{E}[p(\mathbf{f}_*|X, \mathbf{y}, X_*)] = K(X_*, X) \cdot [K(X,X) + \sigma_n^2 I]^{-1} \cdot \mathbf{y} \tag{1.17}$$

$$cov(\overline{\mathbf{f}}_*) = K(X_*, X_*) - K(X_*, X) \cdot [K(X,X) + \sigma_n^2 I]^{-1} \cdot K(X, X_*) \tag{1.18}$$

Per un singolo punto di test f_* con n punti campione la distribuzione di probabilità diventa:

$$p(f_*|X_*, X, \mathbf{f}) \sim \mathcal{N}(\,\overline{f}_*,\ cov(\overline{f}_*)\,) \tag{1.19}$$

$$\overline{f}_* \triangleq \mathbb{E}[p(f_*|X, \mathbf{y}, \mathbf{x}_*)] = K(\mathbf{x}_*, X) \cdot [K(X,X) + \sigma_n^2 I]^{-1} \cdot \mathbf{y} \tag{1.20}$$

$$cov(\overline{\mathbf{f}}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, X) \cdot [K(X,X) + \sigma_n^2 I]^{-1} \cdot K(X, \mathbf{x}_*) \tag{1.21}$$

Essendo la distribuzione di probabilità Gaussiana, il valore più probabile (e dunque la nostra predizione) è la media, mentre la varianza ci dà una misura dell'incertezza della predizione nel punto $\mathbf{x}_*$.

Qualora al posto di considerare processi gaussiani a media nulla volessimo specificare una funzione deterministica $m(\mathbf{x})$, dovremmo semplicemente (per ogni coppia $(\mathbf{x}_{*i}, y_{i*})$ dei punti di test) considerare ogni singola uscita y_{*i} come lo scostamento dalla media al posto che come scostamento dallo zero. La covarianza non cambia, mentre la menda diventa:

$$\overline{\mathbf{f}}_* = m(\mathbf{x}) + K(X_*, X) \cdot [K(X,X) + \sigma_n^2 I]^{-1} \cdot (\mathbf{y} - m(\mathbf{x})) \tag{1.22}$$

1.5.2 La funzione errore

Generalmente, quando effettuiamo una predizione, il valore scelto è $\overline{f}_*$, cioè il valore più probabile, a volte però il valore più probabile non è la scelta

migliore. Pensiamo ad esempio ad un modello relativo a un sensore posto in una centrale nucleare, che misuri in qualche modo il livello di pericolo di esplosione: una sovrastima può portare ad un falso allarme, privo di gravi conseguenze; una sottostima può invece portare ad un incidente con gravi conseguenze. Possiamo quindi definire una funzione errore che quantifichi l'errore che si commette scegliendo il valore $\widetilde{y}$ anziché il valore reale $\widehat{y}$:

$$\mathcal{L}(\widetilde{y}, \widehat{y}) \tag{1.23}$$

Dato un ingresso $\mathbf{x}_*$, l'errore minimo viene commesso scegliendo il valore reale $\widehat{y}$, che però non conosciamo. Ad ogni possibile scelta y_* è associata una probabilità $p(y_*|\mathbf{x}_*, X)$ e un errore $\mathcal{L}(y_*, \widehat{y})$. L'aspettazione dell'errore che compiamo scegliendo $\widetilde{y}$ è quindi:

$$\mathcal{E}(\widetilde{y}|\mathbf{x}_*) = \int \mathcal{L}(\widetilde{y}, y_*) \cdot p(y_*|\mathbf{x}_*, X, \mathbf{y})\, dy_* \tag{1.24}$$

La scelta sarà quindi il valore $\check{y}$ che minimizza l'aspettazione dell'errore:

$$\check{y}_{|\mathbf{x}_*} = \underset{\widetilde{y}}{argmin}[\, \mathcal{E}(\widetilde{y}|\mathbf{x}_*)\,] \tag{1.25}$$

1.5.3 Corrispondenze con il modello a basi di funzioni

Nel caso del modello lineare con basi di funzioni abbiamo $f(\mathbf{x}) = \phi(\mathbf{x}^T)\mathbf{w}$. Supponendo che $\mathbf{w} = \mathcal{N}(\mathbf{0}, \Sigma_p)$, abbiamo:

$$media:$$
$$\mathbb{E}[f(\mathbf{x})] = \phi^T(\mathbf{x})\mathbb{E}[\mathbf{w}] = 0$$
$$covarianza:$$
$$\mathbb{E}[f(\mathbf{x}) \cdot f(\mathbf{x}')] = \phi^T(\mathbf{x})\mathbb{E}[\mathbf{w}, \mathbf{w}^T]\phi(\mathbf{x}') = \phi^T(\mathbf{x})\Sigma_p\phi(\mathbf{x}')$$

Identificando con $k(\mathbf{x}, \mathbf{x}') = \phi^T(\mathbf{x})\Sigma_p\phi(\mathbf{x}')$ abbiamo che $K(X, X') = \Phi^T(X)\Sigma_p\Phi(X')$, quindi le equazioni di media (1.20) e covarianza (1.21) corrispondono con la media e la varianza della distribuzione di probabilità (1.11).

1.5.4 Esempio

Iniziamo con un modello di funzione (unidimensionale) con distribuzione a priori Gaussiana, a media nulla e con funzione covarianza

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 e^{\frac{1}{2}\frac{|\mathbf{x}-\mathbf{x}'|^2}{l}} \tag{1.26}$$

(cioè una funzione covarianza Gaussiana, con $\sigma_f^2 = 1 = l$). In (a) possiamo vedere alcune realizzazioni di questo processo aleatorio nell'intervallo [-10, 10] (linee colorate) e la distribuzione del processo gaussiano a priori (la media è rappresentata dalla linea spessa nera, la regione grigia rappresenta la media$\pm\sigma_f^2$, cioè la regione di confidenza[6]). I grafici "continui" sono stati ottenuti interpolando linearmente $n = 200$ punti equi spaziati.

Prendiamo ora una di queste realizzazioni e campioniamola in 10 punti casuali, ottenendo i dati $\{(x_1, f(x_1)), ..., (x_{10}, f(x_{10}))\}$; supponendo poi che la campionatura sia avvenuta in presenza di un rumore additivo, gaussiano, a media nulla e variabili indipendenti ($\epsilon(0, \sigma_n^2)$) aggiungiamo ad ogni punto campione uno scostamento ottenendo il training set $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ con $y_i = f(x_i) + \epsilon$ (b).

Conoscendo questi 10 punti possiamo calcolare la distribuzione (1.16) sui 200 punti di test la cui media (non avendo specificato alcuna funzione d'errore) rappresenta il valore predetto (c). Nel riquadro (d) mostriamo l'errore in relazione ai dati: nei pressi dei punti campione abbiamo un errore molto basso (dell'ordine del rumore), mentre allontanandosi dai punti l'errore tende a crescere, riflettendo l'allargamento della regione di confidenza.

Lo stesso procedimento è applicato poi ad un modello di funzione con distribuzione a priori che differisce dalla precedente per $m(x) = \sin(x)$; potrebbe essere ad esempio il caso pratico in cui sappiamo che ad una segnale deterministico $\sin(x)$ è stato aggiunto un segnale incognito a media nulla. Viene applicato lo stesso procedimento ottenendo le figure (e-f-g-h).

[6] $f_*(x, \omega)$ è un processo gaussiano. Fissato un punto di ingresso a, $f(a, \omega)$ è una variabile aleatoria Gaussiana $\sim \mathcal{N}(0, \sigma_f^2)$. La regione di confidenza nel punto a è l'intervallo simmetrico $[-\alpha, \alpha]$ tale che $\int_{-\alpha}^{\alpha} f(a, \omega)\, d\omega = 0,95$.

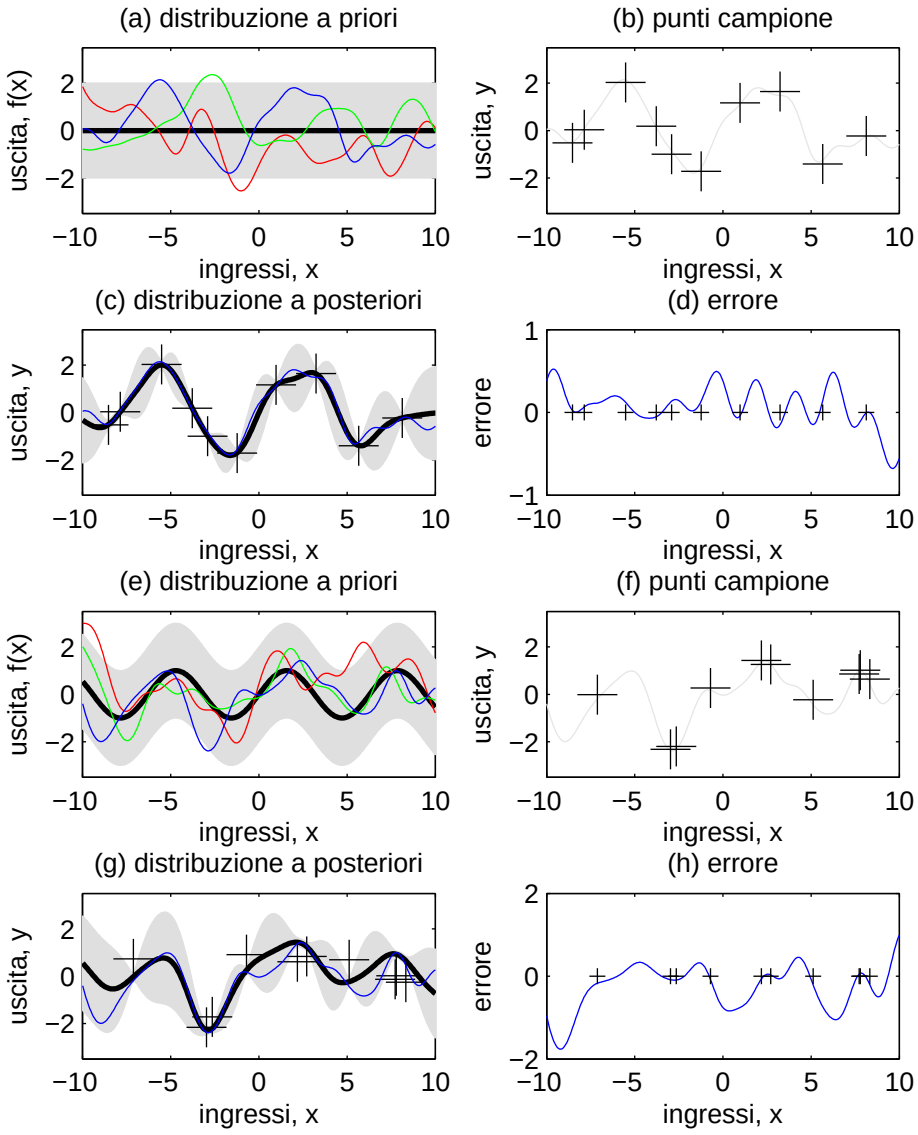


Figura 1.1: Esempio

Capitolo 2

Funzioni covarianza

2.1 Proprietà

Le proprietà di un processo gaussiano dipendono dalla funzione covarianza e dai suoi parametri. Definiamo alcune caratteristiche delle funzioni covarianza.

La funzione covarianza deve essere *simmetrica* $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, in quanto (anche intuitivamente) la correlazione tra due punti non dipende dall'ordine considerato.

Definizioni:

Una funzione covarianza si definisce *stazionaria* se dipende solamente dalla differenza tra i due punti d'ingresso: $k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} - \mathbf{x}')$; in tal caso è necessariamente invariante rispetto alle traslazioni, ma non rispetto alle rotazioni, ed è detta *isotropica.*

Una funziona covarianza si dice *radiale* se dipende solamente dalla distanza tra i due punti d'ingresso: $k(\mathbf{x}, \mathbf{x}') = f(|\mathbf{x} - \mathbf{x}'|) = f(r)$; in tal caso è invariante rispetto alle traslazioni e alle rotazioni.

Una funziona covarianza si dice *dot-product* se dipende solamente dal prodotto scalare tra i due punti d'ingresso: $k(\mathbf{x}, \mathbf{x}') = f(x \cdot x')$. Una funzione *dot-product* è necessariamente invariante rispetto alle rotazioni nell'origine.

Dato un insieme finito di punti $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, definiamo *matrice covarianza* la matrice di Gram della funzione $k(\mathbf{x}, \mathbf{x}')$:

$$\begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}'_1) & k(\mathbf{x}_1, \mathbf{x}'_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}'_n) \\ k(\mathbf{x}_2, \mathbf{x}'_1) & k(\mathbf{x}_2, \mathbf{x}'_2) & & \\ \vdots & & \ddots & \\ k(\mathbf{x}_n, \mathbf{x}'_1) & & & k(\mathbf{x}_n, \mathbf{x}'_n) \end{pmatrix} \tag{2.1}$$

Un processo aleatorio $f(\mathbf{x})$ si definisce continuo in media quadratica nel punto $\mathbf{x}_*$ se, $\forall$ successione $\{\mathbf{x_i}\}_{i\in\mathbb{N}}$ tale che $\mathbf{x_i} \underset{i\longrightarrow+\infty}{\longrightarrow} \mathbf{x_i}$:

$$\lim_{i\to+\infty} \mathbb{E}[|f(\mathbf{x}_i) - f(\mathbf{x}_*)|^2] = 0 \ \wedge \ f(\mathbf{x}_*) < +\infty \tag{2.2}$$

Definiamo la derivata nella i-esima direzione di un processo aleatorio $f(\mathbf{x})$ come:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{h\longrightarrow 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \tag{2.3}$$

dove il limite è il limite in media quadratica.

Un importante parametro della funzione covarianza è la *length-scale* caratteristica. Questo parametro rappresenta la levigatezza della curva:

Teorema:

In un processo aleatorio unidimensionale quasi certamente continuo, Gaussiano, a media zero il numero medio di intersezioni con la retta $y = u$ dal basso verso l'alto in un intervallo unitario è dato da:

$$\mathbb{E}[N_u] = \frac{1}{2\pi}\sqrt{\frac{-k''(0)}{k(0)}} \cdot e^{-\frac{u^2}{2k(0)}} \tag{2.4}$$

Ad esempio, nella funzione covarianza usata nell'esempio del capitolo precedente (1.26) la *length-scale* caratteristica è rappresentata dal parametro l. Variando questo parametro si può rendere la funzione più levigata oppure più aspra: nei grafici in figura 2.1 ripetiamo la generazione delle singole realizzazioni del processo aleatorio con la funzione covarianza (1.26), rispettivamente con $l = 0.1$ e $l = 5$

Teorema:

La somma di due funzioni covarianza è una funziona covarianza. Per verificarlo, consideriamo un processo aleatorio $f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$: la funzione covarianza di $f(\mathbf{x})$ è $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$.

Teorema:

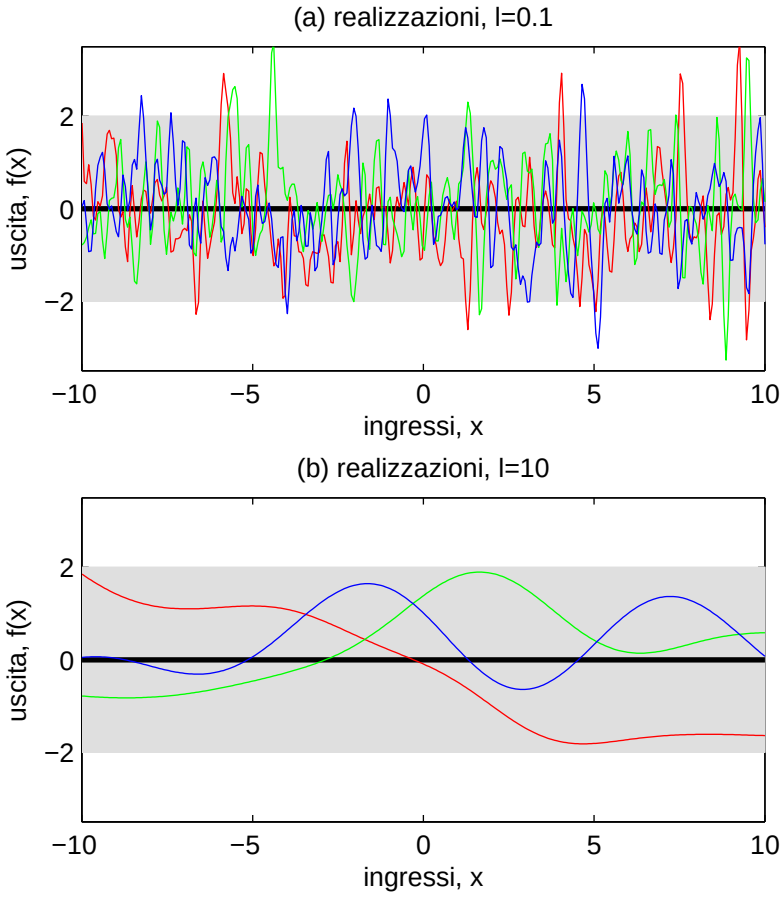


Figura 2.1: Lengthscale

Il prodotto di due funzioni covarianza è una funziona covarianza (la verifica è analoga al caso della somma). Inoltre, se $k(\mathbf{x}, \mathbf{x}')$ è una funzione covarianza valida, allora lo è anche $k^p(\mathbf{x}, \mathbf{x}') \quad \forall p \epsilon \mathbb{N}$

Teorema:

Sia $a(\mathbf{x})$ una funzione deterministica, $f(\mathbf{x})$ un processo aleatorio; consideriamo $g(\mathbf{x}) = a(\mathbf{x}) \cdot f(\mathbf{x})$. Allora $cov(g(\mathbf{x}), g(\mathbf{x}') = a(\mathbf{x}) \cdot k(\mathbf{x}, \mathbf{x}') \cdot a(\mathbf{x}')$. Supponiamo $k(\mathbf{x}, \mathbf{x}') \geq 0 \ \forall (\mathbf{x}, \mathbf{x}')$ e scegliamo $a(\mathbf{x}) = k^{-\frac{1}{2}}(\mathbf{x}, \mathbf{x}')$ da cui $\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x}')} \cdot \sqrt{k(\mathbf{x}, \mathbf{x}')}}$, cioè abbiamo riscalato verticalmente la funzione covarianza, in modo da avere $\tilde{k}(\mathbf{x}, \mathbf{x}') = 1 \ \forall (\mathbf{x}, \mathbf{x}')$.

2.2 Funzione covarianza Gaussiana

La funzione covarianza Gaussiana è la più comunemente usata ed esiste in diverse varianti. Consideriamo dapprima la funzione

$$k_G(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{r^2}{2l^2}} \tag{2.5}$$

dove $r = |\mathbf{x} - \mathbf{x}'|$, dunque la funzione è radiale; l ha il ruolo di length-scale caratteristica: infatti, applicando la 2.4 troviamo che il numero di attraversamenti dell'asse x (in una dimensione) è $(2\pi l)^{-1}$; σ_f è la varianza del processo aleatorio.

Nel caso i dati campione avessero comportamenti molto differenti nelle varie dimensioni, si può arricchire la funzione Gaussiana con differenti length-scale per ogni dimensione:

$$k_G(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T \cdot M \cdot (\mathbf{x}-\mathbf{x}')^T} \text{ dove } M = diag(\mathbf{l})^{-2} \tag{2.6}$$

con $\mathbf{l}$ vettore di D length-scale. In questo modo la funzione covarianza varia più velocemente nelle direzioni degli assi in cui la l corrispondente è più piccola, ciò vuol dire che consideriamo più attendibili (più indicativi, oppure meno rumorosi, ecc) quelle componenti degli ingressi.

Infine, nel caso alcune dimensioni dei dati campione siano correlate tra di loro, potrebbe essere utile specificare alcune direzioni più indicative, diverse dagli assi:

$$k_G(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T \cdot M \cdot (\mathbf{x}-\mathbf{x}')^T} \text{ dove } M = diag(\mathbf{l})^{-2} + \Lambda\Lambda^T \tag{2.7}$$

dove Λ è una matrice $D \times i$ rappresentante gli i vettori delle direzioni desiderate. In tutte le forme presentate la Gaussiana è una funzione stazionaria.

2.3 Cenni su altre funzioni covarianza

Vediamo brevemente altre funzioni covarianza di uso comune.

Matérn La Matérn è una classe di funzioni del tipo:

$$k_{Matern} = \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\frac{r\sqrt{2\nu}}{l} \right)^{\nu} \cdot K_{\nu} \left(\frac{r\sqrt{2\nu}}{l} \right) \text{ con } l, \nu \geq 0 \tag{2.8}$$

con K_ν funzione di Bessel modificata. Per $\nu \to \infty$ ritroviamo la funzione covarianza esponenziale. La funzione è differenziabile in media quadratica n volte, $\forall n < \nu$; i valori di ν più usati sono $\nu = \frac{3}{2}$; $\nu = \frac{5}{2}$, poiché per valori inferiori il processo diventa estremamente variabile, mentre per valori superiori diventa difficilmente distinguibile dalla covarianza Gaussiana.

Esponenziale γ

$$k(r) = e^{-(\frac{r}{l})^{\gamma}} \text{ con } 0 < \gamma \leq 2 \tag{2.9}$$

Questa famiglia comprende anche la covarianza Gaussiana, ed è differenziabile in media quadratica solo per $\gamma = 2$ (nel caso appunto della Gaussiana)

Razionale quadratica

$$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha l^2} \right)^{-\alpha} \text{ con } l, \alpha \geq 0 \tag{2.10}$$

Questa funzione covarianza può essere vista come una somma infinita di covarianza gaussiane, con differenti length-scale caratteristiche. Poniamo $\tau = \frac{1}{l^2}$

$$k_{RQ} = \int \tau^{\alpha-1} e^{-\frac{\alpha\tau}{\beta}} e^{-\frac{\tau r^2}{2}} dr \propto \left(1 + \frac{r^2}{2\alpha l^2} \right)^{-\alpha} \tag{2.11}$$

Polinomiali continue a tratti a supporto compatto Questa classe di funzioni, essendo a supporto compatto, ha la caratteristica che la correlazione tra gli input diventa nulla oltre una certa distanza, introducendo parecchi zeri nella matrice covarianza e dunque rendendo meno onerosa dal punto di vista computazionale l'inversione della matrice. Bisogna però porre attenzione, nel definire questo tipo di funzioni, per garantire che siano definite positive, caratteristica che varia anche a seconda della dimensione dello spazio di ingressi considerato.

Capitolo 3

Scelta del modello e dei parametri

3.1 Approccio gerarchico

Si pone il problema di determinare i parametri, la funzione covarianza e i suoi iperparametri più adatti al modello che stiamo esaminando. L'approccio più comune è dividere il problema in tre livelli gerarchici:

1. parametri $\mathbf{w}$
2. iperparametri θ
3. modello $\mathcal{H}_i$

Si tratta di procedere attraverso questi tre livelli e di determinare, mediante la regola di Bayes, i valori più probabili date le nostre conoscenze sulla funzione. Al primo livello determiniamo la distribuzione a posteriori dei parametri:

$$p(\mathbf{w}|\mathbf{y}, X, \theta, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i) \cdot p(\mathbf{w}|\theta, \mathcal{H}_i)}{p(\mathbf{y}|X, \theta, \mathcal{H}_i)} \tag{3.1}$$

la distribuzione $p(\mathbf{w}|\theta, \mathcal{H}_i)$ è la distribuzione a priori dei parametri, $p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)$ è la verosimiglianza, cioè la probabilità di ottenere i dati campioni posto che i parametri abbiano la distribuzione a priori specificata. La distribuzione al denominatore è la *verosimiglianza marginalizzata* rispetto ai parametri:

$$p(\mathbf{y}|X, \theta, \mathcal{H}_i) = \int p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i) \cdot p(\mathbf{w}|\theta, \mathcal{H}_i)\, d\mathbf{w} \tag{3.2}$$

ed è costante (rispetto a $\mathbf{w}$).

Successivamente, per gli iperparametri θ:

$$p(\theta|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \theta, \mathcal{H}_i) \cdot p(\theta|\mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)} \tag{3.3}$$

Ora, essendo l'incognita i valori θ, $p(\mathbf{y}|X, \theta, \mathcal{H}_i)$ rappresenta la verosimiglianza, cioè la probabilità di ottenere determinati punti campioni posto che gli iperparametri abbiano la distribuzione a priori specificata. Analogamente la costante di normalizzazione rappresenta la verosimiglianza marginalizzata rispetto agli iperparametri:

$$p(\mathbf{y}|X, \mathcal{H}_i) = \int p(\mathbf{y}|X, \theta, \mathcal{H}) \cdot p(\theta|\mathcal{H}_i)\, d\theta \tag{3.4}$$

Infine, applicando ancora una volta la regola di Bayes otteniamo la distribuzione di probabilità a priori del modello $\mathcal{H}_i$:

$$p(\mathcal{H}_i|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathcal{H}) \cdot p(\mathcal{H})}{p(\mathbf{y}|X)} \tag{3.5}$$

ancora una volta la verosimiglianza marginalizzata dell'equazione precedente prende il posto della verosimiglianza e la distribuzione a denominatore rappresenta la marginalizzazione rispetto a $\mathcal{H}_i$ della verosimiglianza (questa volta in forma di sommatoria perché consideriamo i k tipi di funzioni covarianza presi in considerazione:

$$p(\mathbf{y}|X) = \sum_{i}^{k} p(\mathbf{y}|X, \mathcal{H}) \cdot p(\mathcal{H}_i) \tag{3.6}$$

Fissati gli iperparametri θ e la funzione covarianza $\mathcal{H}$, l'equazione 3.1 risulta identica all'equazione trovata per la distribuzione a posteriori 1.5 che portava alla distribuzione dei valori predetti 1.11.

3.2 Verosimiglianza marginalizzata

Purtroppo alcuni di questi integrali, all'atto pratico, risultano complicati da risolvere algebricamente. Esiste un modo alternativo per scegliere gli iperparametri: al posto di scegliere gli iperparametri come il valore più probabile della distribuzione a posteriori (3.3), scegliamo gli iperparametri che massimizzino la verosimiglianza marginalizzata (3.2).

Scegliamo ora una funzione covarianza $\mathcal{H}$ e fissiamo i suoi iperparametri θ. La verosimiglianza marginalizzata è, per definizione:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X) \cdot p(\mathbf{f}|X)\, d\mathbf{f} \tag{3.7}$$

dover la distribuzione a priori $p(\mathbf{f}|x)$ è supposta essere una distribuzione Gaussiana $\mathcal{N}(\mathbf{0}, K)$:

$$p(\mathbf{f}|X) = \frac{1}{\sqrt{(2\pi)^n |K|}} \cdot e^{-\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f}} \tag{3.8}$$

e la verosimiglianza è una distribuzione Gaussiana $\mathcal{N}(\mathbf{f}, \sigma_n^2 I)$:

$$p(\mathbf{y}|\mathbf{f}, X) = \frac{1}{\sqrt{(2\pi\sigma_n^2)^n}} \cdot e^{-\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f}} \tag{3.9}$$

Calcolando l'integrale otteniamo:

$$p(\mathbf{y}|X) = \frac{1}{\sqrt{(2\pi)^2 \cdot |K + \sigma_n^2|}} \cdot e^{-\frac{1}{2}\mathbf{y}^T (K+\sigma_n^2)^{-1}\mathbf{y}} \tag{3.10}$$

di questa quantità si usa negli algoritmi, per comodità, il logaritmo: $\log[p(\mathbf{y}|X)] = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2)^{-1}\mathbf{y} - \frac{1}{2}\log(|K + \sigma_n^2|) - \frac{n}{2}\log 2\pi$,

Dati degli iperparametri θ la verosimiglianza marginalizzata dipende da essi attraverso la matrice covarianza K:

$$\log[p(\mathbf{y}|\theta, X)] = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2)^{-1}\mathbf{y} - \frac{1}{2}\log(|K + \sigma_n^2|) - \frac{n}{2}\log 2\pi \tag{3.11}$$

L'espressione è composta da tre termini. Il primo, cioè l'unico termine in cui compaiono le uscite dei dati campione rappresenta il *data-fit*, cioè quanto i dati sono "credibili" utilizzando una specifica funzione covarianza (e iperparametri). Il secondo, dipendente dalla funzione covarianza (e quindi dagli iperparametri e dal rumore) e rappresenta invece la complessità del modello mentre l'ultimo è solamente una costante di normalizzazione. Per comprendere meglio il ruolo di questi termini, analizziamo l'andamento di questi termini al variare della length-scale nel caso di una funzione covarianza Gaussiana con parametri $l = 1,\ \sigma_f = 1,\ \sigma_n = 0.1$. Nel grafico a) possiamo vedere che il data-fit è una funzione della length-scale monotòna decrescente, in quanto all'aumentare di l il modello diventa sempre meno flessibile, costringendo la funzione predetta a passare più lontana dai dati campione; contestualmente

la complessità diminuisce all'aumentare di l (nel grafico è considerata una *funzione penalità* della complessità, che quindi è monotonamente crescente). La verosimiglianza marginalizzata ha un massimo vicino a $l = 1$ (come ci aspettiamo avendo generato i dati con da una funzione covarianza con $l = 1$) dovuto all'effetto combinato dei due termini, che rappresenta un compromesso tra la complessità del modello e i data-fit (vedi prossimo capitolo). Questo massimo è tanto più evidente quanti più punti campione abbiamo a disposizione (grafico b).

Per massimizzare la verosimiglianza marginalizzata calcoliamo le derivate parziali rispetto ai parametri[1]:

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \theta) = \frac{1}{2}\mathbf{y}^T K^{-1} \cdot \frac{\partial K}{\partial \theta_j} \cdot K^{-1}\mathbf{y} - \frac{1}{2} tr(K^{-1} \cdot \frac{\partial K}{\partial \theta_j})$$

la cui complessità computazionale è $\mathcal{O}(n^3)$ per invertire la matrice K, da calcolarsi una volta sola. Ottenuta K^{-1}, la complessità e $\mathcal{O}(n^2)$.

3.3 Compromesso complessità/data-fit

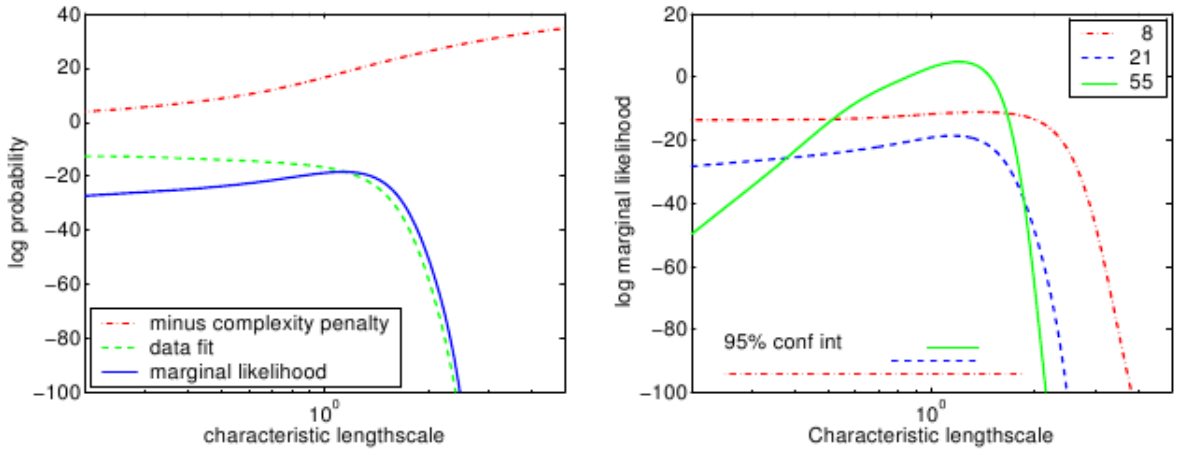
La massimizzazione della verosimiglianza marginalizzata ha il vantaggio di cercare *automaticamente* un compromesso tra la complessità del modello e il data-fit. Un modello semplice (ad esempio un modello con length-scale molto corte) si adatta (cioè "spiega bene") a pochi insiemi di punti campione, mentre un modello più complesso riesce a spiegare meglio altri insiemi di punti campione; di contro, dovendo la distribuzione di probabilità normalizzarsi a uno, un modello semplice avrà probabilità più alte per quei pochi insiemi che riesce a spiegare bene rispetto a un modello complesso. Massimizzare la verosimiglianza marginalizzata vuol dire scegliere la distribuzione che garantisce la probabilità più alta, relativamente all'insieme dei dati campione in considerazione.

[1] Sia K una matrice invertibile. Allora

$$\frac{\partial}{\partial \theta} K^{-1} = -K^{-1} \cdot \frac{\partial}{\partial \theta} K \cdot K^{-1} \tag{3.12}$$

dove le derivate si intendono come le derivate dei singoli elementi delle matrici. Inoltre

$$\frac{\partial}{\partial \theta} \log |K| = tr(K^{-1} \cdot \frac{\partial}{\partial \theta} K) \tag{3.13}$$



(a) (b)

Il concetto è facilmente intuibile dal grafico in figura 3.1 dove in ascissa sono rappresentati tutti i possibili insiemi di dati campione. Relativamente al data-set y_1, la verosimiglianza marginalizzata è massima per la distribuzione più semplice (curva di colore blu), mentre con il data-set y_2 è preferibile la distribuzione che rappresenta un modello più complesso (curva di colore blu). Questo procedimento tende quindi a favorire il modello più semplice tra quelli in grado di spiegare i dati.

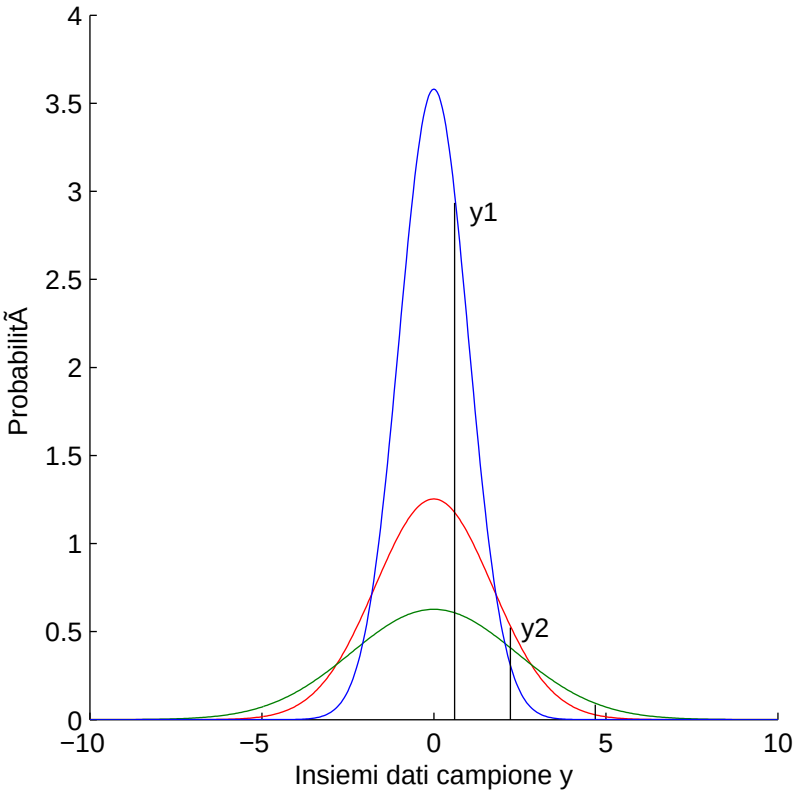


Figura 3.1: Compromesso verosimiglianza marginalizzata

Capitolo 4

Localizzazione

Applichiamo il metodo della regressione Gaussiana ad un caso reale. Lo scopo è localizzare un dispositivo (nodo mobile) che comunica in modalità wireless con dei sensori chiamati nodi ancora.
I sensori sono disposti in maniera uniforme (equidistanti nel piano) sul soffitto di una stanza di sei metri per nove; il nodo mobile comunica con i nodi ancora inviando ininterrottamente pacchetti attraverso un'interfaccia wireless, i nodi ancora sono collegati con una rete USB ad un elaboratore che registra l'*intensità* del segnale con cui arrivano i pacchetti. Ogni dato in ingresso è formato dalle intensità rilevate dagli 8 sensori ed è quindi un punto di uno spazio $\mathbb{R}^8$. Volendo trovare la posizione del dispositivo, i dati in uscita saranno le coppie $\mathbf{z} = (z_x, z_y)$, dunque punti di $\mathbb{R}^2$ (trascuriamo l'altezza da terra, i dati sono stati raccolti tutti con il nodo alla medesima altezza da terra). Per semplicità di trattazione considereremo la posizione lungo x indipendente dalla posizione lungo y: creeremo due modelli separati a partire dai due insiemi di dati campione

$$\mathcal{D}_x = (X, z_x); \quad \mathcal{D}_y = (X, z_y)$$

I modelli che vogliamo costruire sono:

$$z_x = f_x(\mathbf{x}); \quad z_y = f_y(\mathbf{x}) \tag{4.1}$$

4.1 Raccolta dati

I dati campione sono stati raccolti nel laboratorio *NavLab*; i sensori (chiamati nodi ancora) sono stati collegati per mezzo di vari hub USB ad un PC che registrava tutti i dati su un file di testo.

In un primo momento la raccolta dei dati era stata suddivisa in due tipi: a dispositivo mobile fermo e in movimento. Allo scopo di raccogliere agevolmente un grande numero di campioni si era infatti pensato di raccogliere dapprima solo una quantità limitata di campioni a dispositivo fermo, poi spostando a velocità costante lungo una retta il dispositivo, raccogliere un campione ogni secondo. Per far questo, il dispositivo veniva mosso ad una velocità di 10 cm/s circa, mediando sui dati ricevuti in un secondo si ottenevano i dati in ingresso ogni 10 cm di spostamento del sensore.

Questo metodo non ha dato i risultati sperati: i dati raccolti risultavano privi di significato, con inspiegabili variazioni continue e valori decisamente fuori scala. Considerando ad esempio un singolo sensore e muovendo verso di esso il dispositivo mobile, il grafico della potenza registrata dovrebbe essere tendenzialmente crescente, pur avendo oscillazioni dovute a errori e rumore mentre questo non avveniva: i dati non presentavano alcun andamento "preferenziale" oppure le oscillazioni erano talmente elevate da nascondere questo effetto (fig. 4.1) .

Nel secondo tentativo, non capendo le cause che hanno portato ai risultati precedenti, sono stati riprogrammati tutti i sensori e i dati sono stati mediati su un periodo di due secondi, ottenendo un dato ogni 20 cm di spostamento del sensore.

In questo modo la qualità dei dati è migliorata leggermente, ma non in maniera soddisfacente. I dati raccolti con il sensore fermo, invece presentavano molti meno problemi. Facendo alcune prove si è confermato che i picchi e le grandi oscillazioni sono dovute per lo più allo spostamento del sensore: con il sensore immobile, dopo un intervallo di circa un secondo, il segnale tende a stabilizzarsi oscillando di pochi decibel; si è dunque deciso di prendere dati solo a sensore immobile.

Sono stati definiti sette "percorsi" rettilinei di cui quattro paralleli al lato corto della stanza (che chiameremo d'ora in poi orizzontale) e tre lungo il lato lungo (verticale). Il sensore veniva mosso spostandolo velocemente ogni 10 secondi di 20 cm, in modo che i dati abbiano il tempo di stabilizzarsi una volta fermato il sensore: sul grafico ottenuto si notano chiaramente i "picchi"

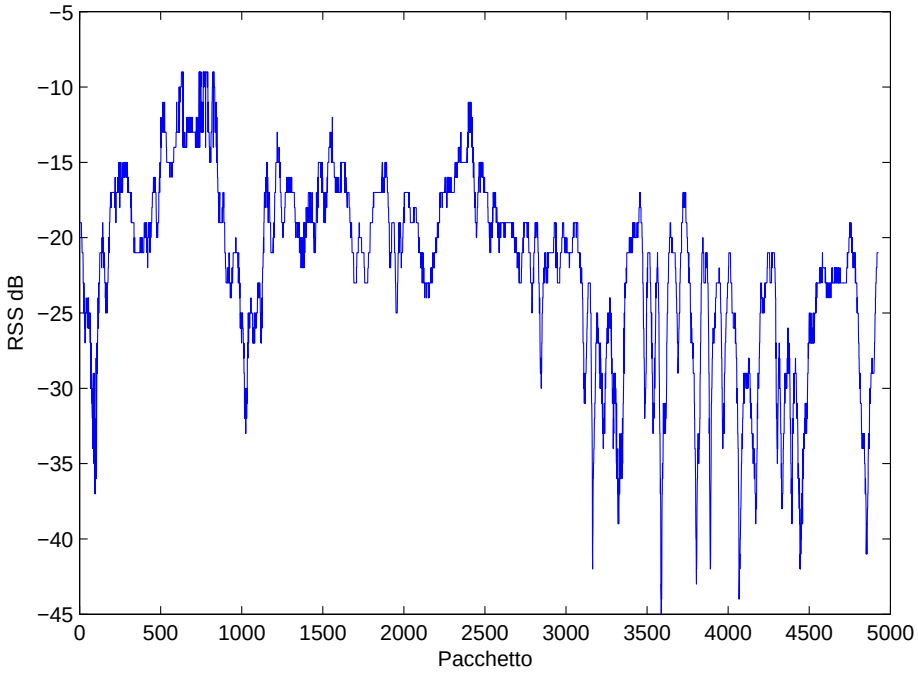


Figura 4.1: Andamento della potenza registrata (RSS) in funzione dei pacchetti ricevuti con nodo mobile in movimento

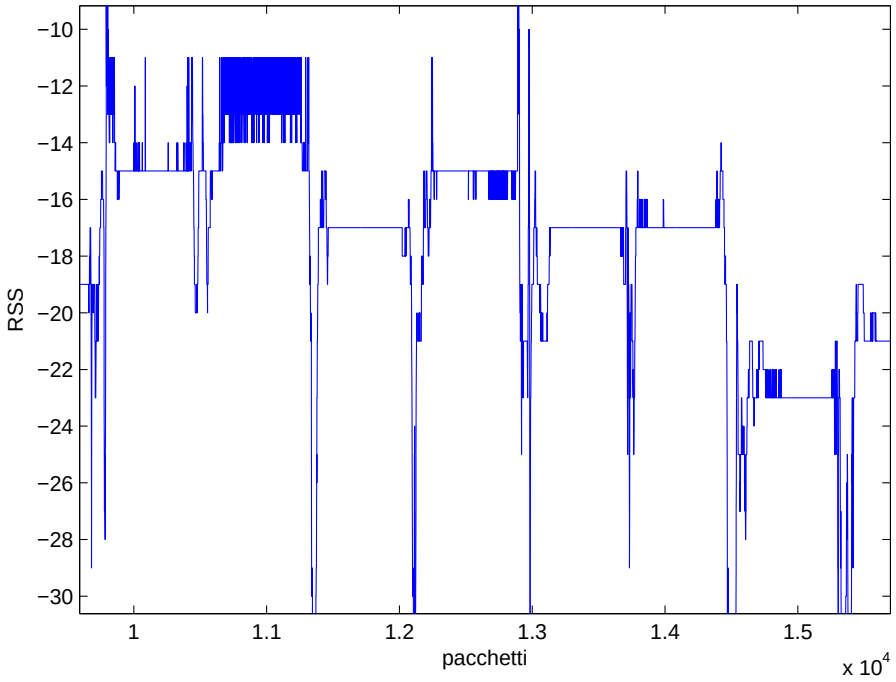


Figura 4.2: Andamento della potenza registrata (RSS) in funzione dei pacchetti ricevuti.

in corrispondenza del movimento del sensore e i periodi più stabili, relativi al sensore immobile (fig. 4.2).

I dati così ottenuti sono stati ulteriormente elaborati: per ognuno dei periodi di 10 secondi sono stati eliminati i 3 secondi iniziali e finali (per eliminare i picchi dovuti al movimento) ed è stata fatta la media sui dati ricevuti nei 4 secondi "centrali", durante i quali il sensore era certamente immobile.

Si noti che comunque l'andamento (relativo ad un sensore verso il quale il dispositivo mobile si stava allontanando) non è monotòno decrescente come ci si aspetterebbe, ma presenta delle oscillazioni anche nel lungo periodo. Questo presumibilmente è dovuto ai fenomeni di riflessione dell'onda elettromagnetica su pareti ed ostacoli; inoltre tavoli, sedie e altri oggetti possono frapporsi fra un sensore il dispositivo mobile in un determinato punto, creando un ostacolo che, invece, nella posizione successiva (venti cm dopo)

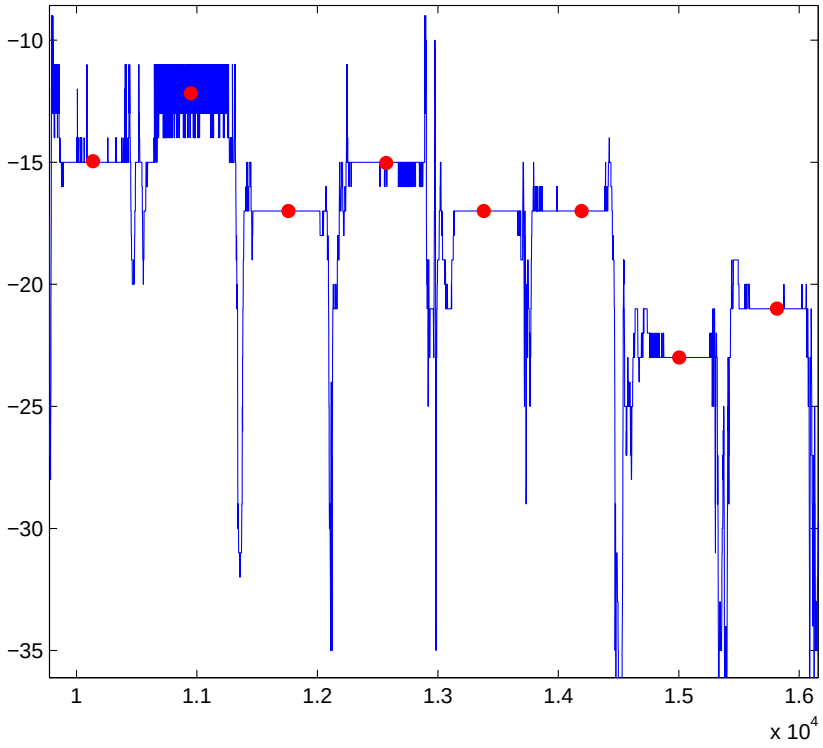


Figura 4.3: Andamento della potenza registrata (RSS) in funzione dei pacchetti ricevuti

non è più presente; nonostante ciò è comunque visibile in tutti i sensori un andamento generale crescente o decrescente a seconda del movimento del dispositivo mobile verso o da un sensore.

I dati raccolti in questo modo sono in numero di 232 campioni.

4.2 Elaborazione

I dati raccolti sono stati organizzati in due matrici separate per creare i due modelli (relativi alla coordinata x e y rappresentanti la posizione del sensore) che vengono elaborati separatamente; considerando i due modelli indipendenti, la distribuzione di probabilità della posizione sul piano si potrà

calcolare semplicemente come il prodotto delle due distribuzioni ed essendo le due distribuzioni gaussiane risulterà anch'essa Gaussiana.

Per creare il modello utilizzeremo il metodo della massimizzazione della verosimiglianza marginalizzata. Il training set e il test set sono ottenuti dall'insieme dei dati campione dividendo a metà le misure; assegnando, alternativamente, un dato al training set un dato al test set in modo da ottenere un insieme di test e di training uniformemente spaziato.

Un passo fondamentale nella costruzione del modello è la scelta della funzione covarianza: useremo la più comune covarianza Gaussiana, chiamiamo $l_1, ..., l_8$ i parametri che rappresentano le length-scale (una per ogni sensore), σ_f il parametro che rappresenta la varianza del segnale e un termine σ_n che rappresenta l'offset dei valori misurati (abbiamo preso come origine del sistema di riferimento un angolo della stanza):

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T \cdot M \cdot (\mathbf{x}-\mathbf{x}')^T} + \sigma_n^2 \tag{4.2}$$

$$M = \begin{pmatrix} l_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & l_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & l_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & l_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & l_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & l_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & l_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & l_8 \end{pmatrix} \tag{4.3}$$

Avendo posto un numero di parametri pari al numero dei sensori, quando andremo a massimizzare la verosimiglianza marginalizzata rispetto ai parametri otterremo un effetto conosciuto nella letteratura inglese come ARD (Automatic Relevance Detection): gli ingressi più rumorosi (e perciò meno significativi) otterranno una length-scale più lunga, influendo così in misura minore sui valori predetti. Inoltre qualora notassimo alcune length-scale esageratamente più lunghe delle altre, i dati portati da quei sensori sono probabilmente estremamente più rumorosi: potremmo procedere eliminando i dati provenienti dai corrispettivi sensori, ricalcolando il modello e confrontando l'errore quadratico medio con il risultato precedente.

Per poter confrontare le length-scale è però è necessario che tutti gli ingressi (del training set) vengano riscalati per avere varianza unitaria e me-

dia nulla; corrispondentemente dovremo scalare gli ingressi del test set nella stessa misura.

Nel calcolo del modello si fa uso (per la parte riguardante la massimizzazione della verosimiglianza marginalizzata), di uno script - creato da C.E. Rasmussen e C.K.I. Williams - che calcola il punto di massimo per successive approssimazioni a partire da dei valori iniziali, questo però non garantisce l'identificazione del massimo nel caso di massimi relativi multipli. Iniziamo la ricerca impostando tutti i parametri a 1:

$$M = I_8;\ \sigma_f = 1;\ \sigma_n = 1; \tag{4.4}$$

per il modello relativo alla dimensione orizzontale, troviamo i valori:

$$\begin{array}{ll} l_1 = 0.56 & l_2 = 1.87 \\ l_3 = 2.67 & l_4 = 1.60 \\ l_5 = 1.47 & l_6 = 1.23 \\ l_7 = 4.29 & l_8 = 0.56 \\ \sigma_f = 1.47 & \sigma_n = 0.36 \end{array} \tag{4.5}$$

mentre per quello relativo alla dimensione verticale, i valori trovati sono

$$\begin{array}{ll} l_1 = 1.39 & l_2 = 1.16 \\ l_3 = 1.95 & l_4 = 3.10 \\ l_5 = 2.66 & l_6 = 0.52 \\ l_7 = 0.91 & l_8 = 3.96 \\ \sigma_f = 2.69 & \sigma_n = 0.22 \end{array} \tag{4.6}$$

Prima di procedere dobbiamo verificare in entrambi i casi di non aver scelto un massimo locale. Per far questo ripetiamo la procedura di massimizzazione della verosimiglianza marginalizzata con dati di partenza diversi da 4.4.

Nel caso del primo modello ho ottenuto gli stessi valori per qualsiasi insieme di parametri iniziale, dunque i valori (4.5) sono i valori che useremo per la generazione dei valori predetti. Confrontando le predizioni con i valori reali troviamo uno scarto quadratico medio di 1,37 metri. Nel caso del secondo modello, ho trovato un secondo punto:

$$\begin{array}{ll} l_1 = 20.04 & l_2 = 20.05 \\ l_3 = 20.05 & l_4 = 20.04 \\ l_5 = 20.04 & l_6 = 20.05 \\ l_7 = 20.04 & l_8 = 20.04 \\ \sigma_f = 2.73 & \sigma_n = 2.72 \end{array} \tag{4.7}$$

aumentando le iterazioni dell'algoritmo, non cambiano sensibilmente i valori trovati, dunque ci troviamo nel caso della presenza di due massimi locali. Per valutare quale dei due insiemi di valori è il più adatto per il nostro modello dobbiamo dare una valutazione dell'errore che commettiamo scegliendo l'uno o l'altro insieme di parametri. Dal momento che siamo interessati solamente ad avere il minor errore possibile non utilizziamo alcuna funzione errore, ma calcoliamo solamente lo scarto quadratico medio generando i valori che rappresentano le nostre predizioni e confrontandoli con le uscite del nostro modello.

Vediamo così che nel primo caso lo scarto quadratico medio è di 2.23 metri, mentre nel secondo caso è di 9,7 metri dunque sceglieremo (senza alcun dubbio) il primo insieme di parametri. Procedendo per tentativi non si riescono a trovare altri massimi locali degni di nota (esiste un altro punto di massimo che porta a length-scale dell'ordine di 10^7 facendo sì che i valori predetti siano tutti uguali alla media. Ovviamente tale modello non ha alcuna utilità).

Vediamo ora le previsioni unificate, viste sul piano della stanza (figura 4.6). La figura rispetta le proporzioni reali della stanza ed è relativa ad un solo percorso (per maggiore chiarezza della figura) I punti blu rappresentano la posizione reale del dispositivo mobile, i punti rossi rappresentano la posizione determinata dal modello.

Dai grafici possiamo notare come le previsioni siano concentrate verso l'interno della stanza: quando il modello si muove verso i bordi della stanza l'errore aumenta. Sembra cioè che il modello si "sbilanci" poco nelle previsioni, cercando di mantenere una stima verso il centro della stanza. Possiamo tentare di rendere il modello più propenso a predire valori lontani dalla media introducendo dei dati campione aggiuntivi: per ogni punto dell'insieme di test $(\mathbf{x}, \mathbf{y})$ aggiungiamo due punti con la stessa uscita, ma al primo sommiamo agli ingressi la varianza degli stessi, al secondo la sottraiamo (figura 4.7). Abbiamo così triplicato i punti del training set, dando inoltre informazioni al modello sull'incertezza delle singole misure.

Ottimizziamo i parametri dei due nuovi modelli con lo stesso procedimento del precedente, cominciando dal modello relativo alla direzione orizzontale. In questo caso si trovano diversi punti di massimo della verosimiglianza marginalizzata: per il primo modello tra i punti di massimo trovati, dà lo scarto quadratico medio minore il punto:

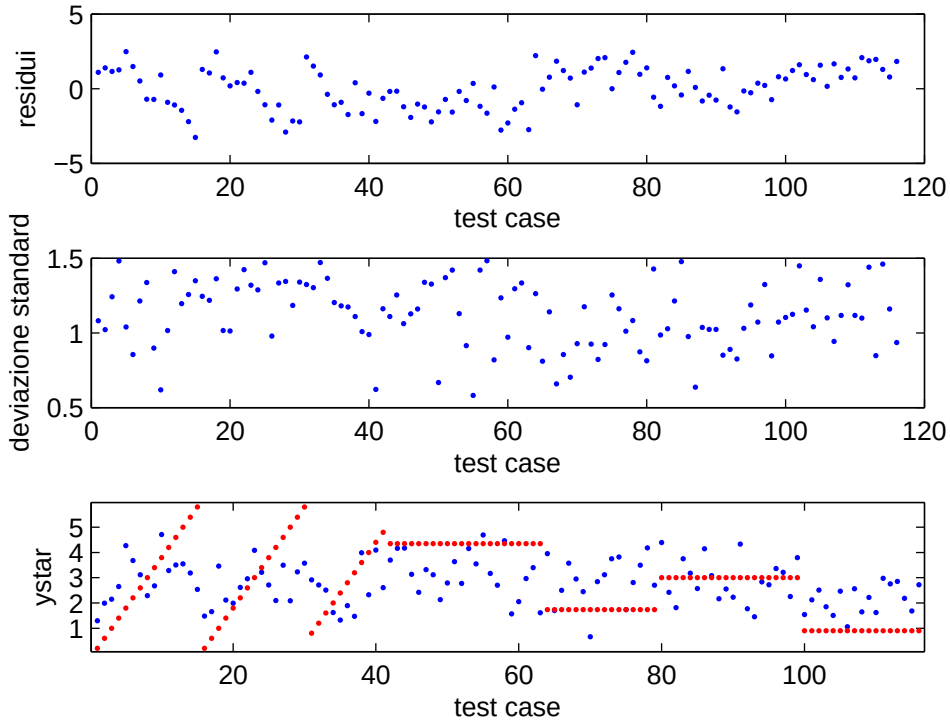


Figura 4.4: Residui, deviazione standard e confronto tra i valori predetti e i valori reali per il modello orizzontale. Nell'ultimo riquadro i punti rossi rappresentano i valori reali della posizione del sensore, i punti blu la predizione.

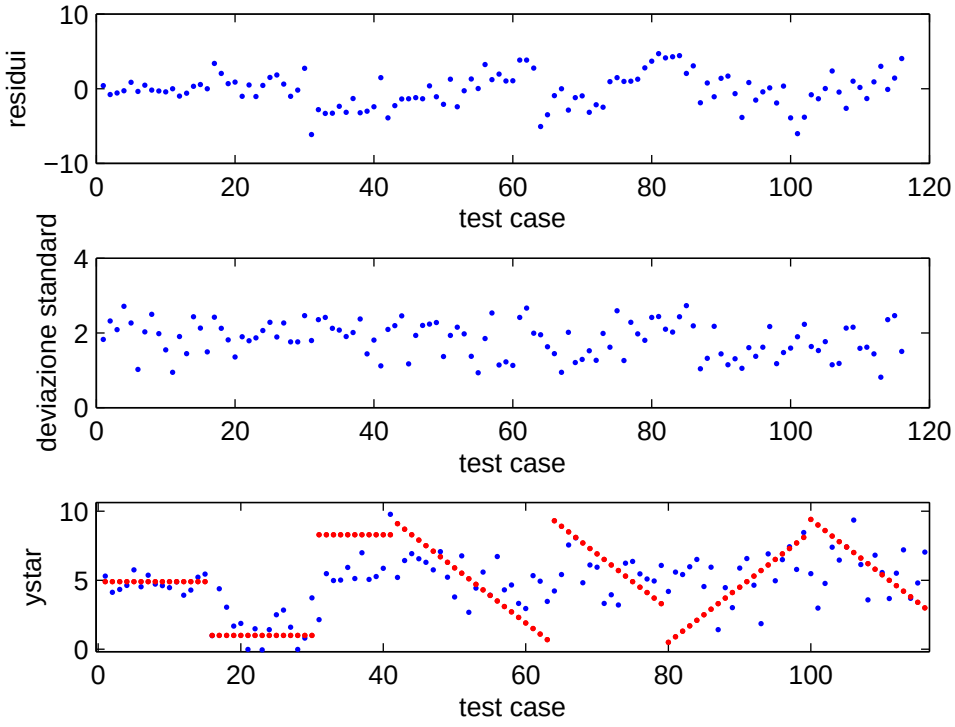


Figura 4.5: Residui, deviazione standard e confronto tra i valori predetti e i valori reali per il modello verticale. Nell'ultimo riquadro i punti rossi rappresentano i valori reali della posizione del sensore, i punti blu la predizione.

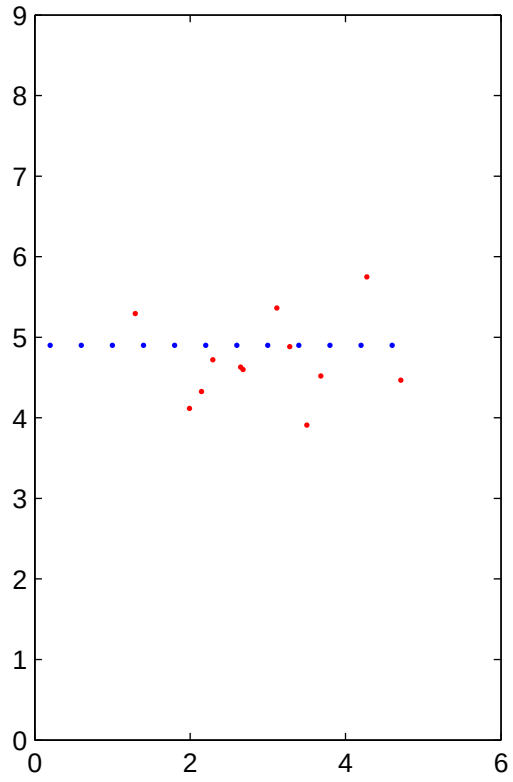


Figura 4.6: Previsioni su un percorso formato da 13 punti

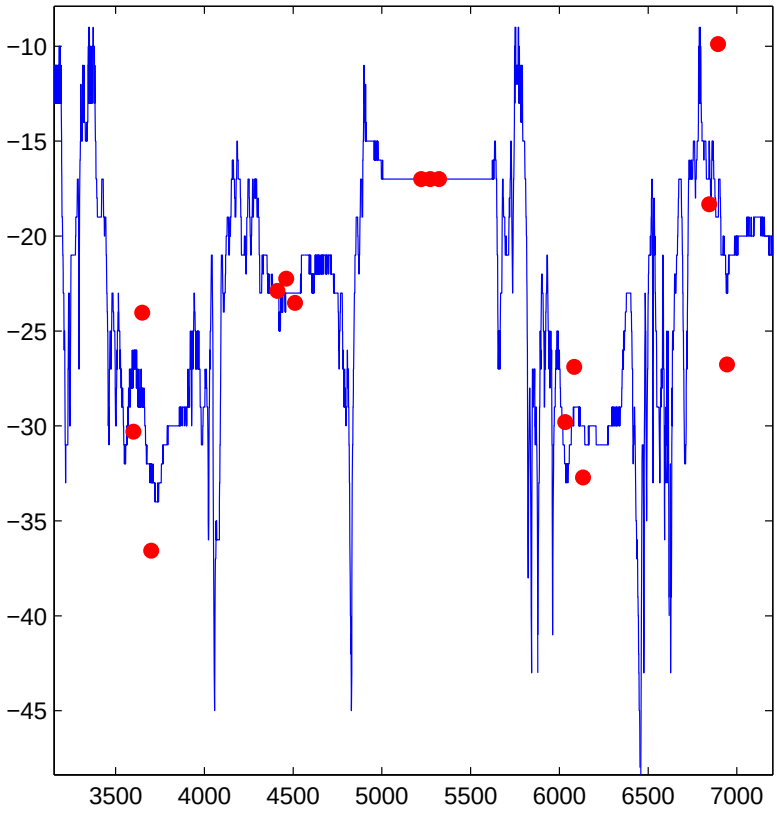


Figura 4.7: Andamento della potenza registrata (RSS) in funzione dei pacchetti ricevuti: per ogni "gruppo" il primo punto rappresenta la media sul periodo (che esclude i picchi transitori), il secondo la media più la varianza e il terzo la media meno la varianza.

$$\begin{array}{ll} l_1 = 2.06 & l_2 = 0.84 \\ l_3 = 0.34 & l_4 = 2.2 \\ l_5 = 1.13 & l_6 = 0.97 \\ l_7 = 0.81 & l_8 = 2.02 \\ \sigma_f = 1.48 & \sigma_n = 0.34 \end{array} \tag{4.8}$$

Lo scarto quadratico medio relativo a questo modello è di 1,41 metri.

Nel caso del modello verticale il punto da preferire è:

$$\begin{array}{ll} l_1 = 1.15 & l_2 = 1.48 \\ l_3 = 0.97 & l_4 = 0.87 \\ l_5 = 1.94 & l_6 = 1.20 \\ l_7 = 0.49 & l_8 = 1.46 \\ \sigma_f = 22.59 & \sigma_n = 0.19 \end{array} \tag{4.9}$$

lo scarto quadratico medio in questo caso è di 2,19 metri. Di nuovo, il fatto che la stanza e i percorsi siano più lunghi sul lato verticale (secondo caso) rende minore la precisione del secondo modello.

Questo nuovo modello ha uno scarto quadratico medio leggermente maggiore nel caso del modello relativo alla direzione orizzontale e leggermente minore nell'altro caso. Questo lieve e parziale miglioramento non giustifica però l'aumento della complessità computazionale (si deve effettuare l'allenamento del sistema con un test set tre volte più numeroso).

Altri tentativi sono stati fatti con l'intento di diminuire la complessità computazionale. Siccome le length-scale (nei casi 4.8 e 4.9) non differivano di molto, ho provato a vedere se era possibile utilizzare un modello più semplice, in modo da dover ottimizzare un numero di iperparametri inferiore, utilizzando una funzione covarianza Gaussiana isometrica:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T \cdot l \cdot (\mathbf{x}-\mathbf{x}')^T} \tag{4.10}$$

che fa uso di tre soli parametri. Inoltre, sempre nell'ottica di rendere più semplice il modello ho rimosso il parametro σ_n, che rappresentava lo scostamento dallo zero dei valori delle uscite, procedendo di conseguenza a centrare l'insieme dei dati in uscita affinché il test set avesse media nulla. La

Infine, notando che i risultati ottenuti con la funzione covarianza con ARD erano migliori, nel tentativo eliminare almeno un iperparametro, ho provato ad utilizzare anche la funzione covarianza Gaussiana:

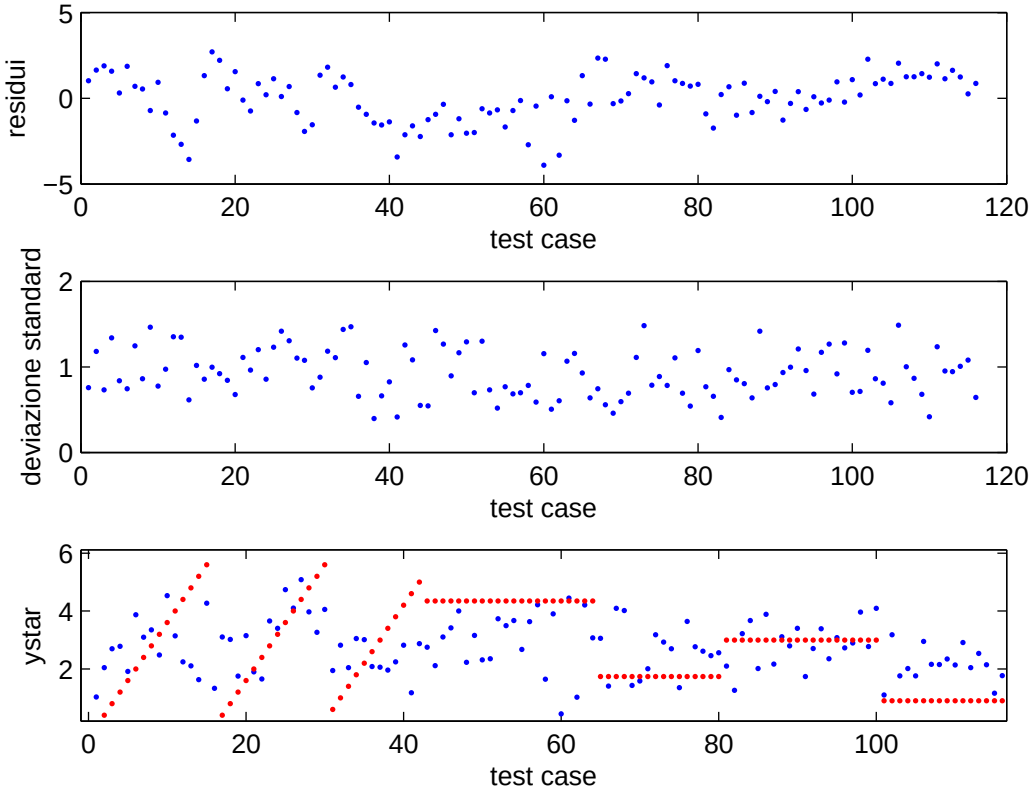


Figura 4.8: Residui, deviazione standard e confronto tra i valori predetti e i valori reali per il modello orizzontale. Nell'ultimo riquadro i punti rossi rappresentano i valori reali della posizione del sensore, i punti blu la predizione.

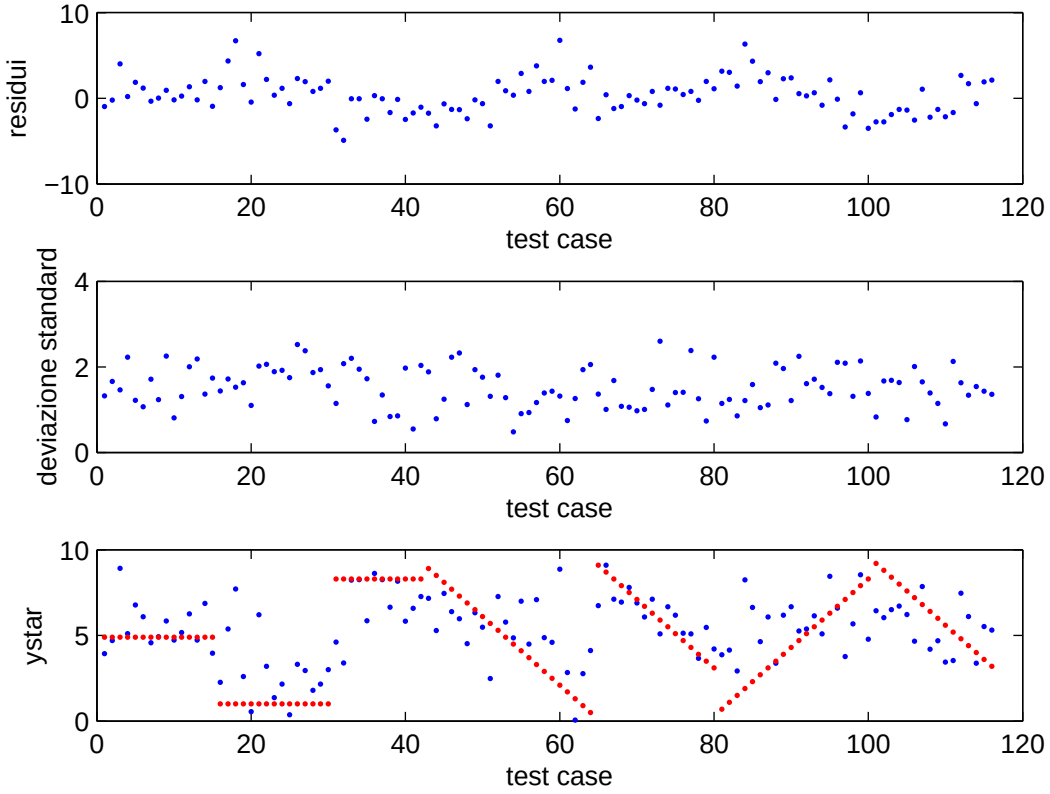


Figura 4.9: Residui, deviazione standard e confronto tra i valori predetti e i valori reali per il modello verticale. Nell'ultimo riquadro i punti rossi rappresentano i valori reali della posizione del sensore, i punti blu la predizione.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T \cdot M \cdot (\mathbf{x}-\mathbf{x}')^T} + \sigma_n^2 \tag{4.11}$$

sempre avendo l'accortezza di centrare le uscite affinché avessero media nulla. Quest'ultima ha ottenuto gli stessi risultati della funzione 4.2.

I risultati migliori sono stati ottenuti con la funzione covarianza Gaussiana non isometrica, cioè con un iperparametro per ogni sensore, ottenendo uno scarto quadratico medio di 2,2 metri.

Capitolo 5

Conclusioni

Generalmente i risultati appaiono in linea con quelli ottenuti precedentemente con modelli parametrici. In particolare, sono soddisfacenti per alcuni insiemi di punti mentre per altri l'errore, se confrontato con la dimensione della stanza in cui sono stati raccolti i dati, assume valori rilevanti.

La differenza tra i vari percorsi è a volte marcata ed è probabilmente dovuta alle differenti condizioni che si verificano nei diversi casi: alcuni percorsi, ad esempio quelli vicini al centro della stanza, passano distante dagli ostacoli quali il mobilio e la strumentazione del laboratorio; in altri percorsi invece questi ultimi si frappongono tra il sensore e il dispositivo mobile impedendo la propagazione diretta dell'onda elettromagnetica.

Un altro elemento di forte disturbo sono anche le parete e altri oggetti di grandi dimensioni che riflettono l'onda elettromagnetica, aumentando ulteriormente l'errore della misura.

Come si può vedere dai grafici del capitolo precedente rappresentanti l'andamento della potenza registrata si può notare quanto questi elementi incidano nella qualità del modello: mentre a fronte di uno spostamento di un passo (20cm) del sensore ci si aspetterebbe una piccola variazione degli ingressi, in certi casi la variazione è di grandezza paragonabile (anche il 25%) a quella tra la prima e l'ultima posizione di un percorso (circa 5 metri).

Con questo tipo di sensori l'approccio della regressione mediante processi Gaussiani potrebbe essere applicato con successo nel caso di ambienti più grandi o poco affollati da ostacoli, come ad esempio negli esterni. Nel caso di ambienti in cui siano presenti ostacoli quali pareti, ecc. la precisione raggiunta è sensibilmente minore ma sufficiente, ad esempio, a determinare (anziché la posizione precisa) delle macro-aree (ad esempio in quale stanza) in cui si

trova il dispositivo, in tal caso potrebbe essere più utile però utilizzare un procedimento di classificazione con processi Gaussiani anziché di regressione.

Bibliografia

- C. E. Rasmussen & C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006
- Federico Girosi, Michael Jones and Tomaso Poggio, *Regularization Theory and Neural Networks Architectures*, 1995
- Radford M. Neal, *Bayesian Learning for Neural Network*s, Springer, 1996