Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

Processi Gaussiani nella Regressione

Relatore:
Chiar.mo Prof.
Andrea Pascucci

Correlatore:
Dott. Pasquale Cascarano

Presentata da:
Vittoria Raulli

Sessione unica
Anno Accademico 2017/2018

Introduzione

L'apprendimento supervisionato è una tecnica di apprendimento automatico (machine learning) che si suddivide in tre passi: nel primo passo vengono raccolti dei dati riguardanti il fenomeno analizzato; nel secondo si costruisce il modello di previsione; nel terzo si applica il modello ottenuto su dei nuovi dati di input.

A seconda delle caratteristiche degli output, l'apprendimento supervisionato si divide in regressione, per output continui, e classificazione, per output discreti. In questa tesi viene presentato il problema della regressione da un punto di vista Bayesiano; tuttavia questo approccio può richiedere costi computazionali molto elevati. Vedremo che i processi Gaussiani risulteranno essere una tecnica molto efficace per risolvere questo problema, sia dal punto di vista computazionale, che dal punto di vista della accuratezza.

La tesi è divisa in 4 capitoli: nel primo vengono presentati i prerequisiti necessari per la lettura; nel secondo viene esposto l'approccio statistico Bayesiano; nel terzo vengono introdotti i processi Gaussiani e nel quarto viene proposto un algoritmo di simulazione.

Indice

Elenco delle figure

Capitolo 1

Prerequisiti

1.1 Distribuzioni Gaussiane

In questa sezione vengono definite le variabili aleatorie con distribuzione Gaussiana e vengono presentati i maggiori risultati su di esse. Per ulteriori approfondimenti, si veda [B].

Definizione 1.1. Si dice che una variabile aleatoria X ha distribuzione normale (o Gaussiana) di media $\mu \in \mathbb{R}$ e varianza $\sigma > 0$ se la sua distribuzione di probabilità è

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}^2\right)\right)$$

Definizione 1.2. Sia $K \in \mathbb{R}^{d\times d}$ simmetrica e definita positiva, sia $\mu \in \mathbb{R}^d$. Si dice che X è una variabile aleatoria con distribuzione Gaussiana multivariata (o distribuzione multinormale) di media μ e matrice di covarianza K se la sua distribuzione di probabilità è

$$\frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{1}{2} < C^{-1}(x-\mu), (x-\mu) >\right)$$

Si osservi che se $d = 1$ le due definizioni coincidono. D'ora in avanti se X è una variabile aleatoria con distribuzione Gaussiana multivariata, scriviamo

$$X \sim \mathcal{N}(\mu, K) \ .$$

1.2 Risultati di analisi numerica

In questa sezione richiamiamo alcuni risultati di analisi numerica che verranno utilizzati nel corso del trattato. Le prove di tali risultati vengono rimandati a [Q].

1.2.1 Il metodo dei minimi quadrati

Siano $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $n > d$, e consideriamo il sistema lineare $Ax = b$. Essendo $n > d$, non è detto che tale sistema abbia soluzioni. Il metodo dei minimi quadrati è una tecnica per calcolare

$$\min_{x\in\mathbb{R}^d} ||b - Ax||_2^2 \ , \tag{1.1}$$

ovvero per approssimare le soluzioni di un sistema sovradeterminato.

Teorema 1.2.1. *Siano $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $n > d$, e sia X l'insieme dei vettori $x \in \mathbb{R}^d$ che risolvono (1.1). Allora valgono le seguenti affermazioni:*

- *$x \in X$ se e solo se è soluzione dell'equazione normale*

$$A^\top Ax = A^\top b \ ;$$

- *X ha un solo elemento se e solo se A ha rango massimo.*

1.2.2 Identità matriciali

Teorema 1.2.2 (Fattorizzazione di Cholesky)**.** *Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica e definita positiva. Alllora esiste una unica matrice L triangolare inferiore tale che $A = LL^\top$.*

Teorema 1.2.3 (Formula di Sherman-Morrison)**.** *Siano $Z \in \mathbb{R}^{n\times n}$ e $W \in \mathbb{R}^{d\times d}$ non singolari, $U, V \in \mathbb{R}^{n\times d}$ tali che $W + V^\top Z^{-1}U \in \mathbb{R}^{d\times d}$ sia non singolare. Allora vale la seguente formula di inversione:*

$$(Z + UWV^\top)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\top Z^{-1}U)^{-1}V^\top Z^{-1} \ .$$

1.3 La regressione lineare standard

Un modello di regressione è un modello matematico che mette in relazione un insieme di variabili indipendenti $x = (X_1, \dots, X_d)$ e una variabile dipendente y, costruito a partire dalla conoscenza di un insieme finito di osservazioni $D = \{(x_i, y_i) | i = 1, \dots, n\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$.

L'obiettivo della regressione è quello di prevedere il valore di y per un nuovo valore di x. In questa sezione verrà analizzato il modello di regressione lineare standard, detto comunemente retta di regressione se $d = 1$, iperpiano di regressione se $d > 1$. Tale modello si presenta nella forma

$$y = f(x, w) = x^\top w.$$

Imponendo le ipotesi di osservazione $y_i = x_i^\top w$, si ottiene un sistema di n equazioni lineari in d incognite $w_1, \dots, w_d$:

$$\begin{cases} y_1 = x_{11} w_1 + \cdots + x_{1d} w_d \\ \vdots \\ y_n = x_{n1} w_1 + \cdots + x_{nd} w_d \end{cases} \tag{1.2}$$

che scriviamo in forma compatta $y = X^\top w$, dove

$$X = \begin{bmatrix} | & \dots & | \\ x_1 & \dots & x_n \\ | & \dots & | \end{bmatrix}, \; y = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, \; w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

Assumendo di avere un numero n di dati molto maggiore di d (cosa che si verifica frequentemente nelle applicazioni), si ha che il sistema (1.2) risulta essere sovradeterminato, e dunque non è detto che esso abbia delle soluzioni. Supponendo $X^\top$ di rango massimo, otteniamo attraverso l'equazione normale

$$XX^\top w = Xy$$

una unica soluzione al problema dei minimi quadrati

$$\min_{w \in \mathbb{R}^d} ||y - X^\top w||_2^2 \ .$$

Tale soluzione $(w_1, \ldots, w_d)$ sarà il vettore dei coefficienti dell'iperpiano di regressione lineare. Un semplice algoritmo per determinare l'unica soluzione $(w_1, \ldots, w_d)$ è considerare la fattorizzazione di Cholesky della matrice $XX^\top$ e risolvere i sistemi

$$Lz = Xy \quad L^\top w = z \ .$$

La fattorizzazione di Cholesky costa $\mathcal{O}(\frac{d^3}{6})$ moltiplicazioni e la risoluzione dei due sistemi triangolari costa $\mathcal{O}(d^2)$ moltiplicazioni, per un costo complessivo di $\mathcal{O}(\frac{d^3}{6}) + \mathcal{O}(d^2) \approx \mathcal{O}(\frac{d^3}{6})$ moltiplicazioni.

Se il numero di dati d è molto grande, la soluzione tramite l'equazione normale risulta essere proibitiva dal punto di vista computazionale. Nel seguente capitolo viene studiato come si sviluppa il modello di regressione lineare da un punto di vista Bayesiano, in cui si suppone di conoscere una probabilità (detta a priori) sui parametri w.

Capitolo 2

Analisi Bayesiana lineare

2.1 Introduzione alla analisi Bayesiana lineare

In questo capitolo viene presentata la regressione lineare Bayesiana, un punto di vista alternativo al metodo dei minimi quadrati per la stima dei parametri di un modello di regressione. L'approccio Bayesiano ha le sue fondamenta in un concetto di probabilità soggettiva, ovvero le probabilità sono interpretate come gradi di fiducia nel verificarsi di un dato evento. Nella inferenza statistica Bayesiana il concetto chiave è il teorema di Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

letto nel modo segente:

- B rappresenta i dati empirici su cui si basa l'inferenza;
- $P(A|B)$ è detta probabilità a posteriori di A, ovvero la probabilità che avvenga l'evento A dopo aver osservato gli eventi B;
- $P(B|A)$ è detta funzione di verosimiglianza;
- $P(A)$ è detta probabilità a priori di A;

- $P(B)$ è detta probabilità marginale, e rappresenta la probabilità di osservare i dati B (matematicamente rappresenta un coefficiente di normalizzazione).

In questa nuovo contesto, il teorema di Bayes può essere riformulato nel segente modo:

$$\text{posteriori} = \frac{\text{verosimiglianza} \times \text{priori}}{\text{verosimiglianza marginale}} \ .$$

2.2 Il modello lineare Bayesiano

Consideriamo l'insieme di osservazioni $D = \{(x_i, y_i)|i = 1, \ldots, n\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, e il modello di regressione lineare $f(x) = x^\top w$. Poichè D consta di dati sperimentali, assumiamo che la variabile dipendente y differisca da $f(x)$ per un rumore di tipo additivo. Si ha dunque

$$y = x^\top w + \varepsilon \ .$$

Assumiamo inoltre che ε sia una variabile aleatoria con distribuzione Gaussiana, di media nulla e varianza σ_n^2:

$$\varepsilon \sim \mathcal{N}(0, \sigma_n^2) \ .$$

Se il numero di dati sperimentali n è molto più grande di d, allora, per quanto detto nel primo capitolo, il calcolo dei parametri w attraverso il metodo dei minimi quadrati risulta molto dispendioso dal punto di vista numerico. Per sopperire al problema del costo computazionale della approssimazione dei parametri w via minimi quadrati, fissiamo una distribuzione di probabilità Gaussiana su ogni parametro w_i $i = 0, \ldots d$ e supponiamo inoltre che ogni w_i sia indipendente da w_j $i \neq j$.

$$w_i \sim \mathcal{N}(0, \Sigma_i) \ .$$

La distribuzione congiunta delle variabili aleatorie $w_1, \dots, w_d$ viene detta **probabilità a priori**. Essendo una variabile aleatoria congiunta di variabili aleatorie gaussiane indipendenti, w risulta avere distribuzione Gaussiana multivariata $w \sim \mathcal{N}(0, \Sigma)$, la cui matrice di covarianza è

$$\Sigma = \begin{bmatrix} Cov(w_1, w_1) & \dots & Cov(w_1, w_d) \\ \vdots & \ddots & \vdots \\ Cov(w_d, w_1) & \dots & Cov(w_d, w_d) \end{bmatrix} = \begin{bmatrix} Cov(w_1, w_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Cov(w_d, w_d) \end{bmatrix}.$$

Si osservi che per l'ipotesi di indipendenza e per le proprietà della covarianza, si ha che Σ è una matrice diagonale, simmetrica e definita positiva.

Definendo

$$X = \begin{bmatrix} | & \dots & | \\ x_1 & \dots & x_n \\ | & \dots & | \end{bmatrix},$$

possiamo adesso ricavare la funzione di verosimiglianza $p(y|X, w)$ con un semplice calcolo:

$$\begin{aligned} p(y|X, w) &= \prod_{i=1}^{n} p(y_i|x_i, w) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\big(-\frac{1}{2}\sigma_n^{-2}(y_i - x_i^\top w)^2\big) \\ &= \frac{1}{(2\pi\sigma_n^2)^{\frac{n}{2}}} \exp\big(-\frac{1}{2}\sigma_n^{-2}||y - X^\top w||^2\big) \\ &\sim \mathcal{N}(X^\top w, \sigma_n^2 I). \end{aligned}$$

Osserviamo che anche la funzione di verosimiglianza risulta essere una distribuzione Gaussiana multivariata.

Applicando il teorema di Bayes è possibile calcolare la probabilità a posteriori sui parametri w. Denotando con $\propto$ la relazione di proporzionalità, si

ottiene:

$$\begin{aligned} p(w|X,y) &\propto p(y|X,w)p(w) \\ &= \exp\big(-\frac{1}{2}\sigma_n^{-2}||y - X^\top w||^2\big)\exp\big(-\frac{1}{2}w^\top\Sigma^{-1}w\big) \\ &= \exp(-\frac{1}{2}\sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w) \end{aligned} \tag{2.1}$$

Lo scopo è quello di massimizzare la probabilità a priori, cioè trovare il $\bar{w}$ che massimizzi la quantità $p(w|X,y)$. Poichè la funzione logaritmo è crescente, allora si ha che massimizzare la quantità $p(w|X,y)$ è equivalente a massimizzarne il logaritmo $\log(p(w|X,y))$. Il problema si riduce quindi alla ricerca di

$$\begin{aligned} &\underset{w}{\text{argmax}}\ (\exp(-\frac{1}{2}(\sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w))) \\ &= \underset{w}{\text{argmax}}\ (\log(\exp(-\frac{1}{2}(\sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w)))) \\ &= \underset{w}{\text{argmax}}\ (-\frac{1}{2}(\sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w)) \\ &= \underset{w}{\text{argmin}}\ (\sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w) \\ &= \underset{w}{\text{argmin}}\ (\psi(w)) \end{aligned}$$

in cui si è posto $\psi(w) = \sigma_n^{-2}||y - X^\top w||^2 + w^\top\Sigma^{-1}w$.
Il $\bar{w}$ cercato, detto **maximum a posteriori** (o più brevemente MAP), è dunque la soluzione dell'equazione

$$\frac{\partial\psi(w)}{\partial w} = 0\ ,$$

la quale, con un semplice calcolo della derivata della funzione $\psi(w)$, si riscrive nella forma

$$-\sigma_n^{-2}X(y - X^\top w) + \Sigma^{-1}w = 0\ .$$

Isolando il termine w si trova la maximum a priori, data da

$$\bar{w} = (\sigma_n^{-2} X X^\top + \Sigma^{-1})^{-1} \sigma_n^{-2} X y = \sigma_n^{-2} A^{-1} X y \ ,$$

in cui si è posto $A = \sigma_n^{-2} X X^\top + \Sigma^{-1}$. Sostituendo $\bar{w}$ in (2.1), si ottiene la **distribuzione a posteriori**

$$p(w|X, y) \sim \mathcal{N}(\bar{w}, A^{-1}) \ ,$$

che risulta anch'essa essere una normale multivariata.

Per predire il valore del modello su un nuovo punto x_* bisogna adesso calcolare la media pesata di tutti i valori che i parametri possono assumere, i cui pesi sono le probabilità che essi vengano assunti. Formalmente, ponendo $f_* = f(x_*)$, si ha che

$$\begin{aligned} p(f_*|x_*, X, y) &= \int p(f_*|x_*, w) p(w|X, y) dw \\ &= \mathcal{N}(\sigma_n^{-2} x_*^\top A^{-1} X y \ , \ x_*^\top A^{-1} x_*) \ . \end{aligned}$$

Si noti che anche la distribuzione predittiva è una Gaussiana, la cui media è la media della distribuzione a posteriori moltiplicata per il vettore di input x_* e la cui varianza è data dalla forma quadratica indotta dalla matrice di covarianza della distribuzione a posteriori A^{-1} calcolata nel punto di input x_*. Da questa osservazione segue che, come ci si aspetta da un modello lineare, il grado di imprecisione del modello cresce in modo proporzionale al numero di dati osservati.

2.3 Un esempio di modello lineare Bayesiano

Mostriamo un esempio di modello lineare Bayesiano nel caso in cui la dimensione dei vettori input è $d = 1$ e si hanno $n = 3$ dati di osservazione. Il modello è della forma

$$y = w_1 + w_2 x \ .$$

Per modelli lineari Bayesiani nel caso $d = 1$, si denota con intercetta (**intercept**) il parametro w_1 e con pendenza (**slope**) il parametro w_2.
Consideriamo la distribuzione a priori sui parametri (w_1, w_2) data da

$$w \sim \mathcal{N}(0, I)$$

come è mostrato nella seguente figura:

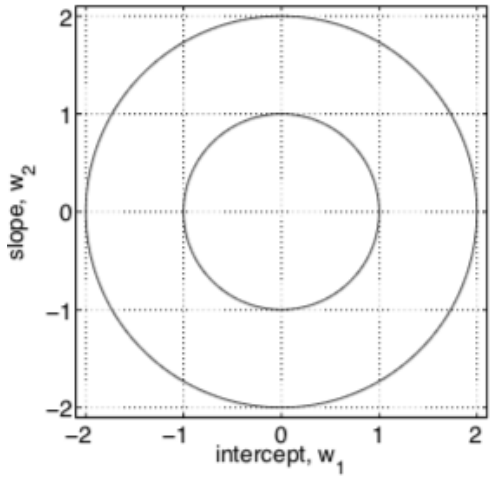

Figura 2.1: Distribuzione a priori[1]

in cui la circonferenza esterna rappresenta il contorno della distribuzione w e la circonferenza interna rappresenta l'area in cui si concentra maggiormente la distribuzione.

Consideriamo 3 punti di osservazione e supponiamo che ci sia un rumore

Gaussiano di tipo additivo dato da

$$\varepsilon \sim \mathcal{N}(0,1) \ .$$

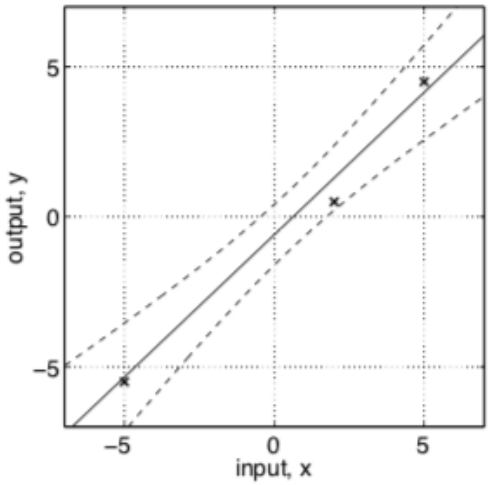

Figura 2.2: Punti di osservazione[1]

Utilizzando le formule ricavate nella sezione precedente, si ottengono la funzione di verosimiglianza e la distribuzione a posteriori.

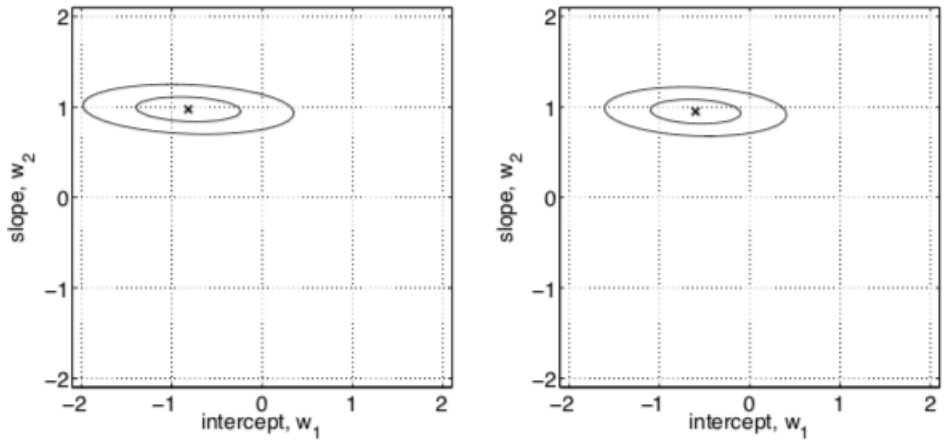

Figura 2.3: Funzione di verosimiglianza (sx) e distribuzione a posteriori (dx)[1]

Nella figura si confronta la funzione di verosimiglianza (a sinistra) con la distribuzione a posteriori (a destra). Si nota come in questo caso il parametro di pendenza sia stato ben approssimato dalla funzione di verosimiglianza, al contrario del parametro di intercetta; questo esempio dunque mostra che per una buona approssimazione non basta la funzione di verosimiglianza.

Da quanto si evince dal grafico della distribuzione a priori, si ha che la maximum a posteriori è

$$(\bar{w_1}, \bar{w_2}) \approx (-\frac{1}{2}, 1) \ .$$

Nella figura 2.2 vediamo come la retta di regressione

$$y = \bar{w_1} + y\bar{w_2}$$

ottenuta attraverso l'approccio Bayesiano sia una buona approssimazione dei dati osservati.

[1]Immagini tratte da: C.E. Rasmussen & C.K.I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, pag. 10

2.4 Regressione in feature space

Il modello di regressione lineare standard $y = x^\top w$ risulta essere il più semplice modello di regressione, in quanto lineare sia rispetto agli input x che ai parametri w. Risulta tuttavia molto limitativo supporre che si possa approssimare qualsiasi fenomeno attraverso un modello lineare rispetto agli input x. Introduciamo quindi un nuovo modello di regresione, in cui si perde la linearità rispetto agli input x, ma si conserva la linearità rispetto ai parametri w. A tale scopo si applica il modello non direttamente sugli input x, ma in un particolare spazio N-dimensionale $\mathbb{S}^N$, detto in letteratura **feature space**. Supponiamo che lo spazio in cui vivono gli input sia d-dimensionale e consideriamo la funzione

$$\begin{aligned}\phi :&\mathbb{R}^d \longrightarrow \mathbb{S}^N \\ &x \mapsto \phi(x) = (\phi_1(x), \dots \phi_N(x))\end{aligned}$$

per delle opportune funzioni $\{\phi_i\}_{i=1,\dots,N}$, che chiamiamo base di funzioni del feature space. Il nuovo modello di regressione nel feature space è

$$y = \phi(x)^\top w$$

in cui il vettore dei parametri w vive in uno spazio N-dimensionale. L'analisi della previsione è analoga al caso della sezione precedente (in particolare ne è una generalizzazione). Consideriamo dunque una distribuzione di probabilità a priori di tipo Gaussiano sui parametri w; analogamente al caso lineare standard, si ottiene la distribuzione predittiva del modello

$$p(f_*|x*, X, y) = \mathcal{N}(\sigma_n^{-2}\phi(x_*)^\top A^{-1}\Phi y \ , \ \phi(x_*)^\top A^{-1}\phi(x_*)) \qquad (2.2)$$

con

$$\Phi = \Phi(X) = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_1(x_n) \\ \vdots & \ddots & \vdots \\ \phi_N(x_1) & \dots & \phi_N(x_n) \end{bmatrix}$$

e $A = \sigma_n^{-2}\Phi\Phi^\top + \Sigma^{-1}$.

Osserviamo che la conoscenza di $p(f_*|x*, X, y)$ comporta il calcolo dell'inversa della matrice $A \in \mathbb{R}^{N\times N}$; questo risulta eccessivamente costoso dal punto di vista computazionale quando N è molto grande. Per sopperire a questo problema, è possibile riscrivere la distribuzione Gaussiana in (2.2) nel seguente modo:

$$\mathcal{N}\big(\phi_*^\top\Sigma\Phi(K+\sigma_n^2 I)^{-1}y \ , \ \phi_*^\top(\Sigma - \Sigma\Phi(K+\sigma_n^2 I)^{-1}\Phi^\top\Sigma)\phi_*\big) \qquad (2.3)$$

con $\phi_* = \phi_(x_*)$ e $K = \Phi^\top\Sigma\Phi$.

Mostriamo infatti che queste distribuzioni normali sono uguali:

- Uguaglianza delle medie:

$$\begin{aligned}
\sigma_n^{-2}\Phi(K+\sigma_n^2 I) &= \sigma_n^{-2}\Phi(\Phi^\top\Sigma\Phi + \sigma_n^2 I) \\
&= \sigma_n^{-2}(\Phi\Phi^\top\Sigma\Phi + \sigma_n^2 I\Phi) \\
&= \sigma_n^{-2}(\Phi\Phi^\top\Sigma + \sigma_n^2 I)\Phi \\
&= \sigma_n^{-2}(\Phi\Phi^\top\Sigma + \sigma_n^2\Sigma^{-1}\Sigma)\Phi \\
&= \sigma_n^{-2}(\Phi\Phi^\top + \sigma_n^2\Sigma^{-1})\Sigma\Phi \\
&= A\Sigma\Phi
\end{aligned}$$

  da cui si ottiene che

$$\sigma_n^{-2}A^{-1}\Phi = \Sigma\Phi(K+\sigma_n^{-2}I)^{-1} \ .$$

- Uguaglianza delle covarianze:
  Segue dalla formula di Sherman-Morrison (Teorema (1.2.3)) ponendo $Z^{-1} = \Sigma$, $W^{-1} = \sigma_n^2 I$, e $V = U = \Phi$.

L'espressione della distribuzione normale in (2.3) necessita del calcolo dell'inversa di una matrice $n\times n$; questo risulta essere molto più conveniente rispetto a quella in (2.2) quando $n << N$ (e ciò si verifica spesso nelle applicazioni).

Capitolo 3

Processi Gaussiani nella regressione

3.1 Processi Gaussiani

Definizione 3.1. Un processo Gaussiano è un processo stocastico $\{f(x)\}_{x\in X}$ tale che, prendendone un qualsiasi numero finito di variabili aleatorie $(f(x_1), \ldots, f(x_n))$, esse hanno una distribuzione di probabilità congiunta Gaussiana.

In un processo gaussiano si ha $f(x_i) \sim \mathcal{N}(\mu_i, \sigma_i)\ \forall i \in I$; pertanto un processo Gaussiano è completamente identificato dalla sua funzione media

$$\begin{aligned} m : X &\longrightarrow \mathbb{R} \\ x &\mapsto \mathbb{E}[f(x)] \end{aligned}$$

e dalla sua funzione di covarianza

$$\begin{aligned} k : X \times X &\longrightarrow \mathbb{R} \\ (x, x') &\mapsto Cov(f(x), f(x'))\ . \end{aligned}$$

Nel seguito, se $f(x)$ è un processo Gaussiano con funzione media m e

funzione di covarianza k, scriveremo

$$f(x) \sim \mathcal{GP}(m, k) \ .$$

Il modello lineare Bayesiano $f(x) = \phi(x)^\top w$ con probabilità a priori $w \sim \mathcal{N}(0, \Sigma)$, studiato nel capitolo precedente, è un esempio di processo Gaussiano. Le sue funzioni media e covarianza sono:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] = \phi(x)^\top \mathbb{E}[w] = 0 \ , \\ k(x, x') &= \phi(x)^\top \mathbb{E}[ww^\top]\phi(x') = \phi(x)^\top \Sigma \phi(x') \ . \end{aligned}$$

Da ciò segue che la distribuzione congiunta delle variabili aleatorie $f(x_1), \ldots, f(x_n)$ è una Gaussiana multivariata di media nulla e covarianza $\phi(x)^\top \Sigma \phi(x')$. La scelta della funzione ϕ specifica dunque una funzione covarianza, nel modo appena descritto. Viceversa, si può mostrare che, fissata una funzione $k(x, x')$ simmetrica e semi definita positiva, esiste una funzione ϕ tale che la funzione covarianza specificata da ϕ è k; tale funzione ϕ può talvolta essere composta da un insieme infinito di funzioni base del feature space. Nel seguito considereremo la funzione di covarianza quadratico-esponenziale (squared exponential):

$$Cov(f(x), f(x')) = k(x, x') = \exp\Big(-\frac{1}{2}||x - x'||^2\Big) \ .$$

Si noti che la covarianza tra gli output è scritta come funzione degli input. Inoltre si osservi che se $x \approx x'$, allora $k(x, x') \approx 1$; mentre se $||x - x'||^2 >> 0$, si ha che $k(x, x') << 1$. Si può mostrare che la scelta di una funzione covarianza specifica una distribuzione sulle funzioni; in particolare la funzione di covarianza squared exponential rende il processo infinitamente differenziabile.

3.2 Predizione senza rumore

Consideriamo il caso in cui le osservazioni sperimentali non siano affette da alcun tipo di rumore. L' ipotesi dell'assenza di rumore è ragionevole in alcuni casi, per esempio nelle simulazioni al computer.
Sia $\{(x_i, f_i)|i = 1, \ldots, n\}$ l'insieme delle osservazioni, e sia X_* l'insieme dei nuovi punti test. La distribuzione congiunta a priori delle uscite delle osservazioni con le uscite test è data da:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) .$$

Per ottenere la distribuzione a posteriori sulle funzioni bisogna restringere la distribuzione congiunta a priori sulle funzioni consistenti con dati di osservazione. Da un punto di vista probabilistico, questo significa condizionare la distribuzione a priori rispetto alle osservazioni. Matematicamente:

$$f_*|X_*, X, f \sim \mathcal{N}(\mu, C)$$

con

- $\mu = K(X_*, X)K(X, X)^{-1}f$;
- $C = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)$.

3.3 Predizione con rumore

Consideriamo adesso il caso in cui le osservazioni siano affette dal rumore $\varepsilon \sim (0, \sigma_n^2 I)$. Il modello è dato da:

$$y = f(x) + \varepsilon .$$

Si ottiene dunque la distribuzione a priori sulle osservazioni:

$$cov(y_p, y_q) = k(x_p, x_q) + \delta_{pq}\sigma_n^2 ,$$

(in cui δ_{pq} è la delta di Kronecker) da cui si ricava la distribuzione congiunta a priori delle uscite delle osservazioni con le uscite test:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) .$$

In modo analogo al caso del modello senza rumore, si ottiene che la distribuzione a posteriori è

$$f_*|X, y, X_* \sim \mathcal{N}(\bar{f}_* \ , \ Cov(f_*)) \ ,$$

in cui

- $\bar{f}_* = K(X_*,X)(K(X,X) + \sigma_n^2 I)^{-1} y$;
- $Cov(f_*) = K(X_*,X_*) - K(X_*,X)(K(X,X) + \sigma_n^2 I)^{-1} K(X,X_*)$.

Se si considera un solo punto di test x_*, l'equazione predittiva è della forma

$$\begin{aligned} \bar{f}_* &= k_*^\top (K + \sigma_n^2 I)^{-1} y \ , \\ \mathbb{V}[f_*] &= k(x_*,x_*) - k_*^\top (K + \sigma_n^2 I)^{-1} k_* \ . \end{aligned}$$

Si osservi che $\bar{f}_*$ si scrive come combinazione lineare delle osservazioni y. Inoltre questa equazione può essere riscritta nella forma

$$\bar{f}_*(x_*) = \sum_{i=1}^{n} \alpha_i k(x_i, x_*)$$

con $\alpha = (K + \sigma_n^2 I)^{-1} y$, ovvero come combinazione lineare di n funzioni covarianza, ognuna di esse centrata in un punto di osservazione. Questa nuova scrittura di $\bar{f}_*$ dice che il processo Gaussiano definisce una distribuzione Gaussiana congiunta sulle variabili y_i, ognuna associata ad un punto dell'insieme X. La base di funzioni può essere infinita; tuttavia per fare una predizione sul punto x_* è sufficiente considerare la distribuzione $n+1$-dimensionale definita dal punto x_* di test e dalle n osservazioni. Si può inoltre notare un'altra

proprietà rilevante del processo Gaussiano: la funzione covarianza $\mathbb{V}[f_*]$ non dipende dai target osservati ma solo dagli input. Infine la distribuzione predittiva y_* sul nuovo punto di test x_* si calcola aggiungendo il termine di rumore.

Capitolo 4

Implementazione della regressione con processi Gaussiani

In questo capitolo è proposto un algoritmo di simulazione della predizione per processi Gaussiani, facendo riferimento ai risultati teorici ottenuti nella sezione 3.3. L'algoritmo è stato implementato in Python 2.7.

4.1 L'algoritmo

Nel codice vengono definite le funzioni:

- `ExpKernel`, che definisce la funzione di covarianza squared exponential;
- `sampleGP`, che costruisce la distribuzione a priori $f(x) \sim \mathcal{GP}(0, k(X, X))$;
- `posteriorGP`, la quale, dato l'insieme X_*, produce una simulazione della predizione $\bar{f}_*$.

Per la simulazione si è usata la funzione `np.arange` per definire il dominio X del processo Gaussiano; in questo esempio $X = \{-5, -4.9, -4.8, \ldots, 4.8, 4.9, 5\}$. Come punti di osservazione sono stati presi $(-2, -2), (-1, -1), (1, 0), (2, 1), (4, -3)$,

che nei grafici sono evidenziati con un asterisco.

L'algoritmo è il seguente:

```
def ExpKernel(x, y):
    cov = np.exp(-0.5*np.power((x.T-y),2))
    return cov

def sampleGP(x, nSamples):
    cov = ExpKernel(x,x)
    f = np.zeros([nSamples, x.shape[1]])
    mean = np.zeros(x.shape[1])
    for i in range(nSamples):
        f[i,:] = np.random.multivariate_normal(mean, cov)
    return f

def posteriorGP(x_star, x, y, nSamples, sigma):
    var = np.eye(x.shape[1])*sigma**2
    inverse = np.linalg.inv(ExpKernel(x,x)+var)
    mat1 = np.dot(ExpKernel(x_star, x),inverse)
    cov = ExpKernel(x_star,x_star)-np.dot(mat1,ExpKernel(x,x_star))
    mean = np.dot(mat1,y.T)[:,0]
    for i in range(nSamples):
        f[i,:] = np.random.multivariate_normal(mean, cov)
    return f

x = np.arange(-5,5,0.1)[np.newaxis]
f = sampleGP(x,5)

for i in range(0,5):
    plt.plot(f[i,:])
plt.title("Prior sampling")
plt.show()

x = np.array([-2,-1,1,2,4])[np.newaxis]
y = np.array([-2,-1,0,1,-3])[np.newaxis]
x_star = np.arange(-5,5,0.1)[np.newaxis]
f_star = posteriorGP(x_star, x, y, 5, 0.2)
plt.plot(x,y, '*')
for i in range(0,5):
    plt.plot(x_star[0,:], f_star[i,:])

plt.title("Posterior sampling")
plt.show()
```

4.2 Analisi dei risultati

Possiamo notare come la distribuzione a priori generata dall'algoritmo in modo casuale (Figura 4.1) non sia un buon modello di predizione.

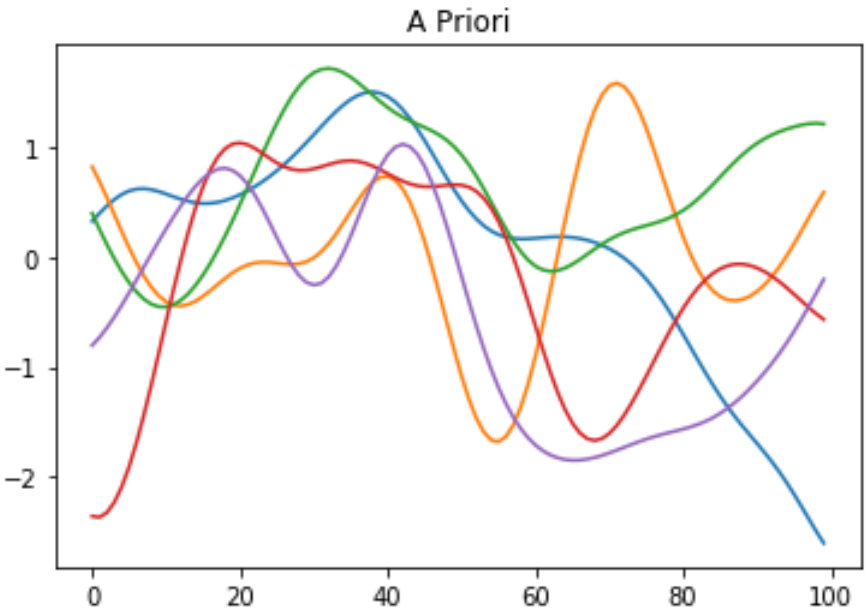

Figura 4.1: GP a priori

Osserviamo invece, come ci si aspetta dai risultati teorici visti nel capitolo precedente, che la distribuzione a posteriori, calcolata dall'algoritmo, genera delle funzioni di approssimazione lisce e coerenti con i dati di osservazione forniti in input. Infatti si vede che l'approssimazione senza rumore interpola i punti di osservazione (Figura 4.3), mentre l'approssimazione con rumore (Figura 4.2) risulta comunque molto efficiente.

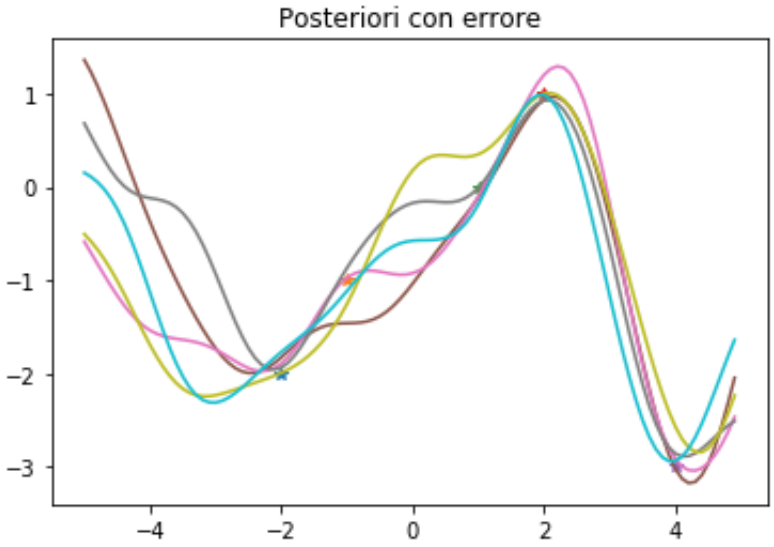

Figura 4.2: Distribuzione a posteriori con errore

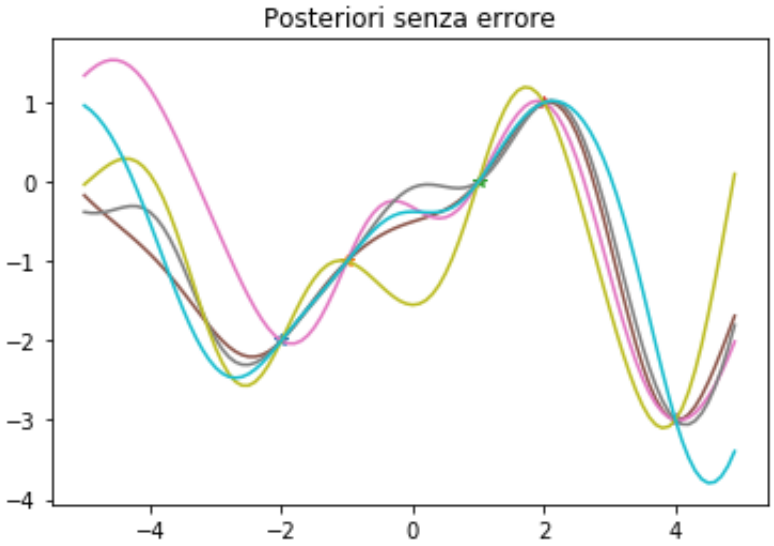

Figura 4.3: Distribuzione a posteriori senza errore

Bibliografia

[B] Pattern Recognition and Machine Learning, C.M. Bishop, Springer.

[Q] Matematica numerica, A. Quarteroni, R. Sacco, F. Saleri, Springer, 2^a edizione.

[R] C.E. Rasmussen & C.K.I. Williams, Gaussian Processes for Machine Learning, the MIT Press.