Riduzione della dimensione di un dataset.

EM 17/18

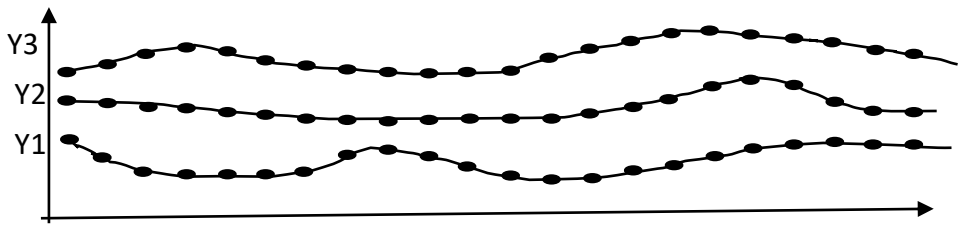
1. La tesina

l problema affrontato in questa tesina è il prodotto di matrici bidimensionali in modalità mutithread. Realizzare una funzione multithread in C in ambiente Linux per fare prodotti fra matrici. Utilizzare la funzione per risolvere un problema di riduzione della dimensione di un dataset.
Testare l'algoritmo su un dataset che verrà fornito in un secondo tempo.

2. Cenni di un algoritmo per la riduzione della dimensione di un dataset

Come accennato nel punto 1, il problema applicativo è di descrivere un evento fisico con il minor numero di dati in maniera ottimale. I motivi più popolari possono essere ad esempio di comprimere l'insieme dei dati (data compression) o di cercare una qualche struttura nei dati (data mining). Una struttura dei dati è visibile solo se le dimensioni sono limitate. Gli algoritmi di riduzione della dimensionalità sono fondamentali nel Data Mining.

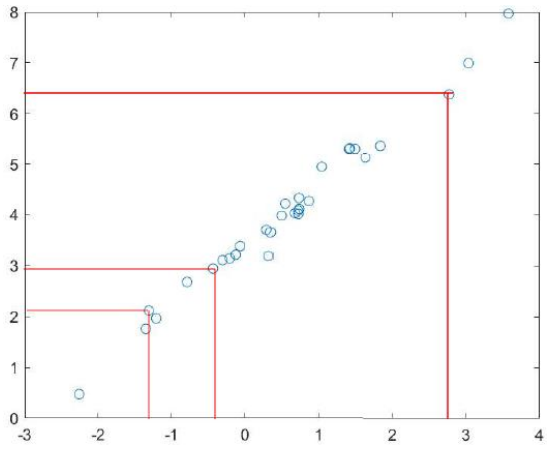
Un evento fisico può essere descritto da una serie di valori, che può essere rappresentata con un grafico od una tabella. Ad esempio le tre curve della seguente figura, dopo un campionamento rappresentato dai punti scuri, possono essere descritte con la tabella bidimensionale sottostante



X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
Y1	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
Y2	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
Y3	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o

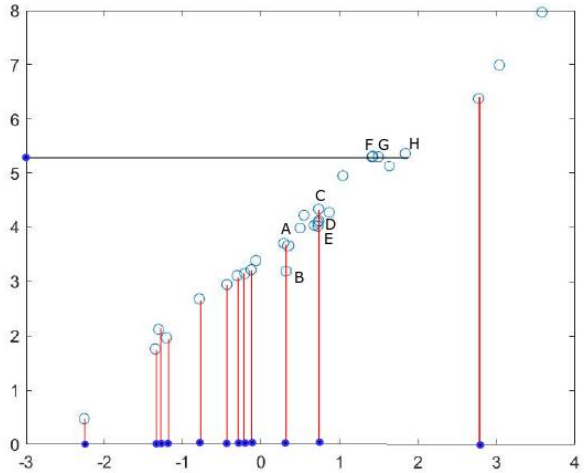
dove per semplicità al posto dei valori numerici è stata scritta una 'o'. In questa tabella naturalmente ogni riga della matrice è rappresentata come un punto in un sistema di coordinate i cui assi sono definiti dalle colonne cioè X-Y1-Y2-Y3.
In questo caso abbiamo quattro variabili da misurare, X, Y1, Y2, Y3. Ad esempio il tempo (rappresentato da X) e le coordinate (x,y,z) di un punto che si muove in uno spazio 3D, dove x è rappresentata da Y1, y da Y2 e z da Y3. Naturalmente, all'aumentare delle variabili misurate aumenta il numero di coordinate. Nel sistema di riferimento X-Y1-Y2-Y3 un punto è un vettore a 4 dimensioni.
Lo scopo è di rappresentare queste misure con un numero di variabili inferiore, considerando solo le variabili con maggiore variazione e scartando quelle che restano relativamente invariate.

In altre parole, quello che si vuol fare è di costruisce un nuovo spazio su cui rappresentare i dati ridefinendo gli assi utilizzando direzioni di maggiore varianza delle variabili originali. I nuovi assi sono chiamati Principal Components da cui il termine Principal Components Analysis o PCA.
Come proiettare i dati originali sul nuovo sistema di riferimento? Considerando un esempio grafico, vediamo un insieme di dati in uno spazio 2D. Le coordinate (x,y) dei dati indicati con cerchietti blu sono ovviamente individuate dalle righe rosse, riportate solo per qualche punto per semplicità



Cosa vuol dire trovare una struttura nei dati? In questo caso può voler dire banalmente che i dati sono allungati secondo la bisettrice ma queste considerazioni sono impossibile se le dimensioni fossero molte di più.

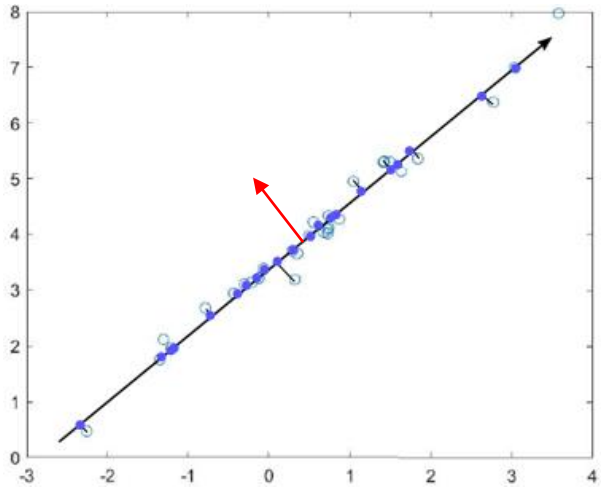
Se i punti vengono proiettati sulle ascisse, otteniamo un nuovo set di punti individuati dai cerchietti blu. Analogamente proiettando sulle ordinate. In entrambi i casi abbiamo una riduzione della dimensione con conseguente riduzione delle coordinate (data compression) ma la proiezione comporta un errore importante. Per esempio, non è più possibile separare i punti quando sono molto sovrapposti. Quando i punti vengono proiettati sulle ascisse non è possibile distinguere i punti A,B e i punti F,G,H. Quando vengono proiettati sulle ordinate non è possibile distinguere tra F, G, H. Oltre a questo problema c'è un problema di distanza tra i punti che viene distorta. Nel senso che nello spazio 2D ci sono dei punti abbastanza distanti che nella proiezione sembrano vicissimi.



Anche in questo caso per semplicità sono considerati solo alcuni punti.

Come si determinano gli assi per i quali questo errore di riduzione della dimensionalità è minore?

Il primo asse evidentemente è la direzione lungo cui si registra la massima dispersione dei dati. Il secondo descrive la rimanente dispersione. Questi assi sono le Componenti Principali del dataset. In questo caso la prima componente principale è l'asse visualizzato nella seguente figura



Proiettando i dati su quest'asse si produce il minimo errore di proiezione. Osservare che secondo la 1° Componente Principale i dati sono maggiormente dispersi. Le Componenti Principali sono mutualmente ortogonali. La seconda Componente è quindi l'asse riportato in rosso nella figura precedente e si vede ad occhio che i dati sono meno dispersi secondo questa direzione.

Le componenti principali successive spiegano una sempre minore percentuale della variabilità originale.
Seguendo questo principio è possibile dire che le ultime componenti principali descrivono principalmente "rumore" ovvero il contributo degli errori di misura o informazioni irrilevanti.

Come si determinano le componenti principali? Il punto principale è la determinazione della matrice di covarianza dell'insieme di dati. Sia $\mathbf{X}$ una matrice di dati, ad esempio

$$\mathbf{X} = \begin{bmatrix} 4.0 & 2.0 & 0.6 \\ 4.2 & 2.1 & 0.59 \\ 3.9 & 2.0 & 0.58 \\ 4.3 & 2.1 & 0.62 \\ 4.1 & 2.2 & 0.63 \end{bmatrix}$$

un insieme di 5 variabili caratterizzanti un evento fisico, ognuna misurata tre volte. Ovvero ogni riga **Xi** è un vettore che rappresenta un insieme di tre misure. Possiamo pensare la matrice come i valori che assumono tre variabili (le colonne) nelle tre misure.

Rappresentiamo la matrice esplicitando i numeri delle variabili e delle misure. Ad esempio la variabile var_1 quando viene misurata una prima volta ha il valore x11, una seconda x12 e così via. La variabile var_2 quando viene misurata una prima volta ha il valore x21, una seconda x22 etc. Chiamiamo N il numero di misure ed M il numero di variabili. In sintesi

$X =$

X11	X12	X13	Var1 → μ_1
X21	X22	X23	Var2→ μ_2
X31	X32	X33	Var3→ μ_3
X41	X42	X43	Var4→ μ_4
X51	X52	X53	Var5→ μ_5

1° misura 2° misura 3° misura

La variabile i-esima ha un valor medio chiamato μi. Questi valori medi sono valutati sui valori delle righe, che sono tutti i valori che la variabile può avere. In altre parole la 1° riga ha media $\mu 1$, la seconda ha media $\mu 2$ e così via. La media μi è ovviamente E[x_i], valutata come $\mu i = \Sigma x_{ik}$, k=1..3. La covarianza di due variabili var_i e var_j è definita come $Cov(var_i, var_j) = E[(var_i - \mu i)(var_j - \mu j)]$.
L'elemento i,j della matrice della covarianza Cov viene cioè calcolato nel seguente modo:

$Cov_{ij} = \Sigma(x_{ik} - \mu_i)(x_{jk} - \mu_j)$, $k=1..N$.

Ovviamente i=1..M e j=1..N. Cioè la matrice di covarianza ha dimensioni M x N. Poniamo M=N. Si vede subito che la matrice di covarianza è simmetrica e che gli elementi sulla diagonale, Cov_{11}, Cov_{22},…,Cov_{MM} sono le varianze delle M variabili.

Calcoliamo ora gli autovalori e autovettori della matrice simmetrica della covarianza.

Si può dimostrare che l'autovettore con il più alto autovalore è la direzione lungo la quale il set di dati ha la varianza massima. Iterando questo risultato si conclude che le **T** componenti principali di un insieme di dati si ottengono trovando gli autovalori e autovettori della matrice della covarianza e scegliendo gli autovettori corrispondenti ai maggiori **T** autovalori.

In sintesi, prendendo questi autovettori e usandoli per ricostruire una approssimazione dei dati originali mediante proiezione (dei dati originali sulle componenti principali che sono in numero inferiore alla dimensionalità dei dati) si realizza un tipo di compressione dei dati. Più componenti principali sono considerati nella proiezione, più i nuovi dati assomiglieranno ai dati originali. In realtà dopo un certo numero di componenti principali l'approssimazione comincia a non migliorare più. In ogni caso sono le prime componenti principali che consentono di ricostruire la maggior parte dei dati originali.

Come si proiettano i dati sugli assi definiti dalle Componenti Principali?

Sia C_P la matrice delle Componenti Principali, cioè la matrice degli autovettori corrispondenti ai maggiori **T** autovalori. Chiamiamo X_P i dati proiettati sulle componenti principali. Si può dimostrare che

$X_P = (X - M) * C_P$

Dove l'elemento i,j della matrice M è definito come $M_{ij} = \mu_i$ cioè le righe sono la media della variabile i.

Questi sono i dati originali proiettati sulle componenti principali (quindi ridotti di dimensione) definiti nello spazio delle componenti originali.

Volendo avere i dati nello spazio originale basta invertire l'equazione matriciale:

$X_P * C_P^{-1} = (X - M) \rightarrow X = X_P * C_P^{-1} + M$

L’unica cosa che bisogna controllare è che i prodotti matriciali siano realizzabili se consideriamo **T < N.**

Come si vede l’algoritmo richiede diverse operazioni di algebra lineare, come il calcolo della covarianza, della media, degli autovettori/autovalori e dell’inversione di matrice.
L’operazione di prodotto verrà fatta con il programma multhread del punto 1.
Le operazioni matriciali più complesse verranno fatte usando una libreria di algebra lineare (si veda https://en.wikipedia.org/wiki/Comparison_of_linear_algebra_libraries per sceglierne una).