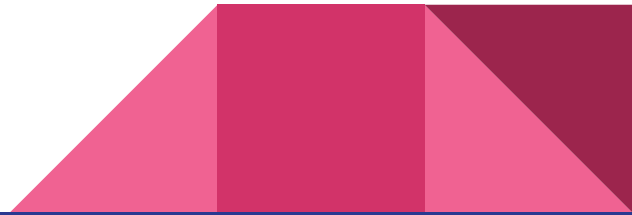
20. Analisi delle componenti principali (PCA)

Corso di Python per il Calcolo Scientifico

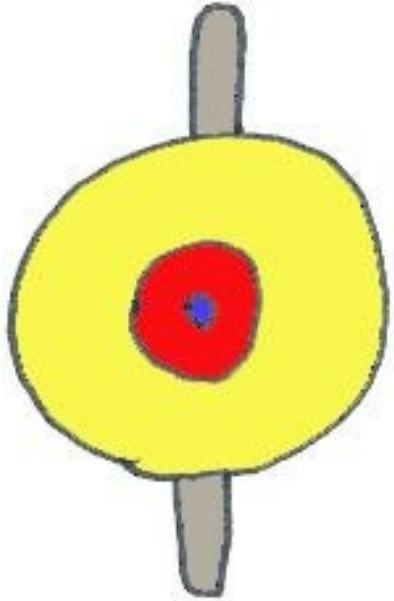
Vito Renò
vito.reno@stiima.cnr.it

Outline

- Definizione e motivazioni
- Esempi pratici
- Appendice: l’algoritmo



Definizione e motivazioni



Che cosa rappresenta ???

Definizione e motivazioni

- Analisi statistica che nasce da un problema fondamentale: **come fare per analizzare una grande quantità di dati espressi utilizzando un numero elevato di variabili?**

- Fatto: Una delle difficoltà insite nella statistica multivariata è il problema di visualizzare dati che hanno molte variabili.
- Osservazione: Nei set di dati con molte variabili, spesso alcuni gruppi di variabili hanno un andamento simile, ovvero è possibile che alcune varibili siano in qualche modo legate l'una con l'altra.

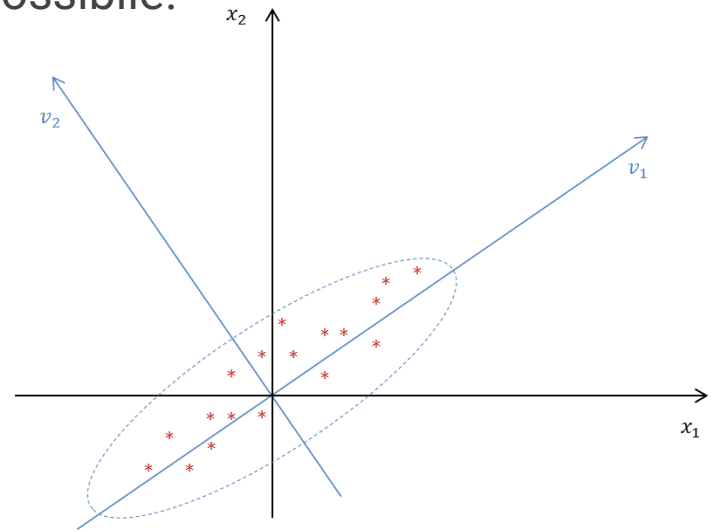
Definizione e motivazioni

- È possibile semplificare il problema sostituendo/cambiando le variabili?

- L'analisi delle componenti principali è un metodo quantitativamente rigoroso per ottenere questa semplificazione mediante la generazione di un **nuovo insieme di variabili**, chiamate componenti principali.
- La PCA permette di compattare la rappresentazione dei dati a disposizione in funzione di opportune variabili artificiali che riescono a visualizzare meglio il dominio da studiare.

Definizione e motivazioni

- Geometricamente, immaginando tutti i dati come una nuvola di punti in uno spazio n-dimensionale, la PCA non fa altro che trovare quel sottospazio vettoriale dove la stessa nuvola, proiettata, appare deformata il meno possibile.



La PCA, nella pratica, corrisponde ad un cambio di riferimento che massimizza l'informazione visibile nei dati esaltando le direzioni lungo cui la varianza è massima

Definizione e motivazioni

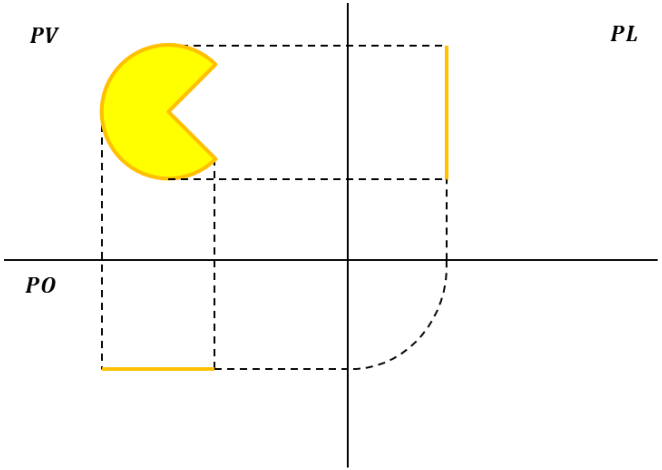
- Ogni componente principale è una **combinazione lineare** delle variabili originali. Tutte le componenti principali sono ortogonali tra loro (quindi non ci sono informazioni ridondanti) e formano una base ortogonale per lo spazio dei dati.

- L'analisi delle componenti principali è un algoritmo **non supervisionato** per la **riduzione della dimensionalità** di un dataset che **identifica** e **scarta** le caratteristiche ritenute **meno utili ad approssimare** il set di dati.

Definizione e motivazioni

Alcuni vantaggi derivanti dall'uso della PCA:

- Contenimento/riduzione del rischio di overfitting se si utilizzano delle feature rumorose
- Velocizzazione della fase di addestramento di un algoritmo di machine learning
- Migliore visualizzazione dei dati (da n-dimensioni a 2 o 3 dimensioni)



La PCA cambia il punto di vista di chi osserva i dati rappresentandoli nel sottospazio definito dalle variabili artificiali, che in questo esempio equivale a guardare il piano verticale della proiezione, ovvero PacMan così come chiunque lo immagina.

Esempi pratici

- Notebook 1: PCA sul dataset IRIS
- Notebook 2: PCA su un dataset di votazioni scolastiche
- Notebook 3: PCA applicata alle immagini: Eigenbackground

Esempi pratici

- Libreria di riferimento

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Appendice: l'algoritmo

Formalmente, a partire da una matrice di N vettori colonna

$$X = [x_1 \quad x_2 \quad \dots \quad x_N]$$

rappresentanti gli N campioni iniziali, si vogliono determinare al più N vettori

$$Y = [y_1 \quad y_2 \quad \dots \quad y_N]$$

detti componenti principali che siano non correlati, ovvero tali che la matrice di covarianza nello spazio trasformato sia diagonale.

$$cov\{Y\} = E\{(Y - E\{Y\})(Y - E\{Y\})^T\} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_N \end{bmatrix}$$

Appendice: l'algoritmo

In particolare, il generico λ_i rappresenta la varianza del vettore y_i

$$var\{y_i\} = \lambda_i \text{ per } i = 1 \dots N$$

che è la massima possibile, mentre si riduce al minimo la correlazione tra due diversi vettori nello spazio trasformato dato che

$$cov\{Y\}_{ij} = 0 \text{ per ogni } i \neq j.$$

I vettori di Y, quindi, sono trasformazioni lineari dei vettori di X che verificano la proprietà su descritta e possono essere espressi come

$$Y = TX$$

Appendice: l'algoritmo

La matrice di trasformazione T si ottiene diagonalizzando la matrice di covarianza di X

$$\Sigma = cov\{X\}$$

Dal momento che Σ è per definizione una matrice quadrata di ordine N simmetrica e semidefinita positiva, essa ammette N autovalori

$$\lambda_1, \lambda_2, \dots, \lambda_N \geq 0$$

ed N autovettori ortonormali

$$e_1, e_2, \dots, e_N$$

che servono a costruire, per righe, la matrice di trasformazione $T = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_N^T \end{bmatrix}$

Appendice: l’algoritmo

Solitamente gli autovalori di Σ vengono scritti in ordine decrescente, così che le componenti principali risultino ordinate rispetto alla varianza, dalla maggiore alla minore.

È possibile analizzare un insieme di dati con l’analisi delle componenti principali anche utilizzando la **decomposizione ai valori singolari**.

Appendice: l'algoritmo

sia A la matrice dei dati in ingresso, organizzata in modo tale che ogni vettore colonna rappresenti le k informazioni relative ad un individuo

$$A = [a_1 \quad a_2 \quad \dots \quad a_N]$$
$$a_i \in \Re^k$$

dalla quale si possono standardizzare le variabili statistiche in ingresso ottenendo una matrice di dati a valore atteso nullo

$$B = A - E[A] = [b_1 \quad b_2 \quad \dots \quad b_N]$$
$$b_i \in \Re^k$$

Calcolando la decomposizione ai valori singolari di $B = USV^T$ si possono trarre le seguenti conclusioni

Appendice: l'algoritmo

La matrice $U \in \Re^{kxk}$ è la matrice ortogonale dei vettori singolari sinistri che corrispondono agli autovettori della matrice BB^T, in questo caso equivalenti a quelli della matrice delle covarianze di A perché

$$BB^T = \big[A - E[A]\big]\big[A - E[A]\big]^T$$

ha gli stessi autovettori di

$$cov\{A\} = E\left\{\big[A - E[A]\big]\big[A - E[A]\big]^T\right\}$$

I vettori singolari sinistri di B, pertanto, non sono altro che le componenti principali di A;

Appendice: l'algoritmo

La matrice $S \in \Re^{NxN}$ è la matrice diagonale dei valori singolari di B, ognuno dei quali è pari alla radice quadrata del corrispondente autovalore della matrice BB^T perché

$$BB^T = USS^TU^T$$

e la matrice SS^T è anch'essa diagonale con gli elementi pari a quelli di S elevati al quadrato. La SVD organizza i valori singolari in ordine decrescente, quindi anche le componenti principali sono già ordinate rispetto alla loro varianza per costruzione.

Domande?

42