Dimension Reduction

Corso AI Engineering - Lezione 3

Reti

INDICE

DIMENSION REDUCTION

- ✓ Dimension Reduction
- ✓ Feature Selection VS Feature Extraction
- ✓ Principal Component Analysis
- ✓ Isomap

Supervised Learning

- Output categorico:
  - **Classificazione**
- Output numerico:
  - **Regressione**

Unsupervised Learning

- Output categorico:
  - **Clustering**
- Output numerico:
  - **Dimension Reduction**

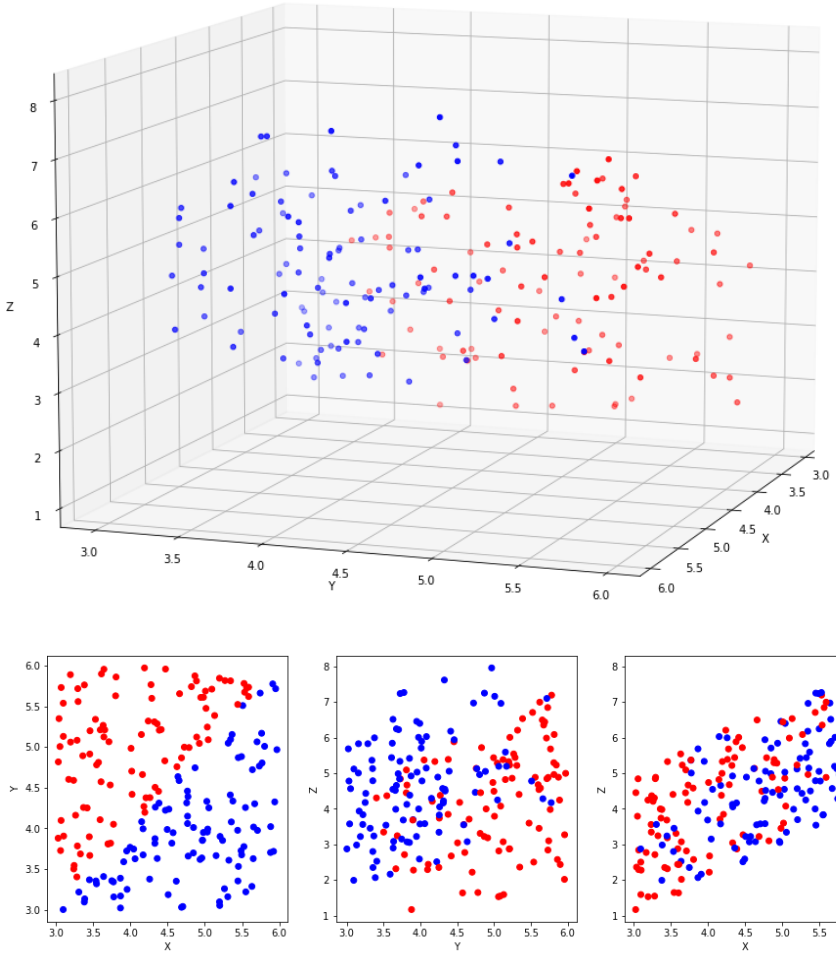
Dimension Reduction

Curse of dimensionality: lavorare con tante dimensioni porta ad avere dati *sparsi,* ovvero dati molto distanti tra di loro perchè lo spazio in cui sono rappresentati ha un volume eccessivo

I dati sparsi mettono in difficoltà modelli che si basano su:

- **significabilità statistica**: poche osservazioni per uno spazio esteso
- **similitudine:** difficile individuare somiglianze in dati non *vicini*

La **Dimension Reduction** consiste in una trasformazione dei dati finalizzata alla riduzione del numero di dimensioni



Selection

La **feature selection** è il processo di selezione di un sottoinsieme di attributi rilevanti per il modello in definizione.

L'ipotesi fondamentale nella selezione è che nel dataset siano **presenti degli attributi che siano rindondanti o irrilevanti** che possano essere rimossi senza troppa perdita di informazione.

Alcuni algoritmi usano delle tecniche finalizzate alla selezione di attributi.

Quali?

Extraction

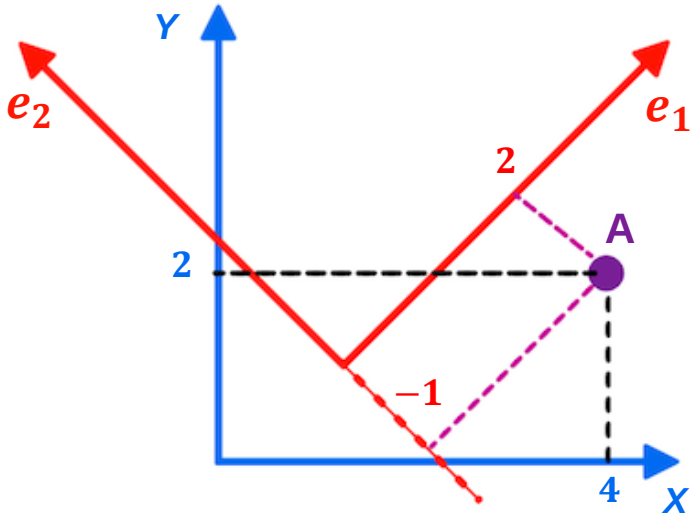
La **feature extraction** costruisce un dataset di feature *derivate* maggiormente informative e non rindondanti, che tuttavia descrivono il dataset con sufficiente accuratezza

Quando un dataset molto largo e c'è il sospetto di dimensioni rindondanti, allora è possible **trasformarlo** in un insieme ridotto di feature, detto *feature vector*.

Una feature presente nel feature vector è una combinazione di più attributi del dataset originale.

Trasformazioni spaziali

- Uno spazio N-dimensionale è descritto con un sistema di N coordinate spaziali. In questo sistema possiamo individuare dei vettori di lunghezza unitaria, o **unit vectors**.
- Un insieme di N unit vector ortogonali creano un nuovo sistema di coordinate. Ogni unit vector rappresenta **la direzione di una componente del dato** rappresentato.
- È possibile passare da un sistema all'altro tramite una **trasformazione lineare ortogonale**. Tale trasformazione **preserva la geometria dello spazio.**



$$A = (4, 2)_{XY}$$

$$A = (2, -1)_{e_1 e_2}$$

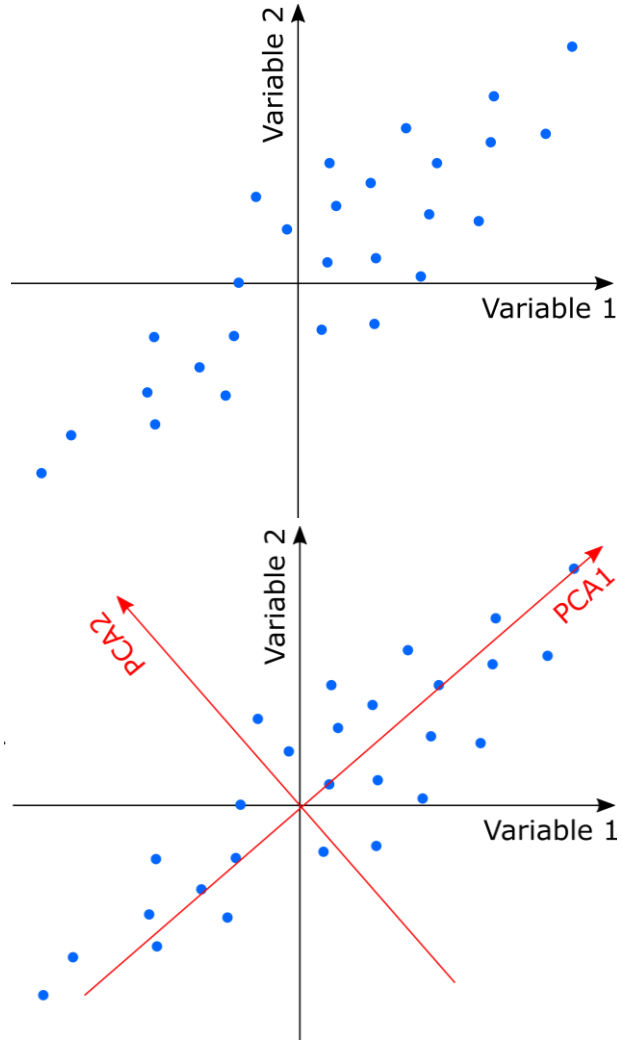
PCA

L'obiettivo della **Principal Component Analysis (PCA)** è definire un *sistema di coordinate gerarchiche* che colgano progressivamente il massimo della varianza del dataset:

- Nel dataset si trova la direzione che minimizza la distanza dai punti trovando la prima direzione principale.
- Con lo stesso criterio si aggiungono altre direzioni ortogonali a quelle già individuate.

Per la natura gerarchica delle dimensioni principali, lo spazio dei dati sarà ben approssimabile usando una rappresentazione con un numero inferiore di features *derivate*

Funziona bene con dati scalati e comportamenti lineari tra le feature



Varietà

Quando un dataset mostra relazioni non-lineari, la PCA non riesce a produrre risultati interessanti.

Tuttavia il dataset potrebbe comunque possedere relazioni lineari in porzioni ristrette di spazio.

Una **varietà** (o **manifold**) è uno spazio che *localmente* ha le proprietà di uno spazio euclideo lineare.

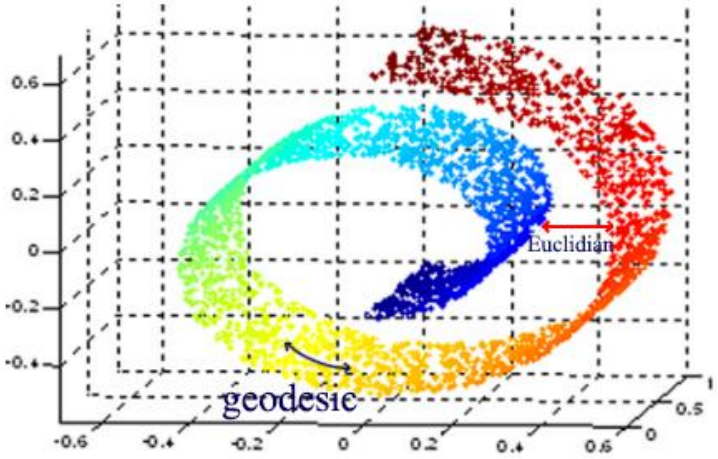
La ricerca di varietà K-dimensionali in uno spazio N-dimensionale (con N>K) può portare a ottimi risultati in termini di riduzione delle dimensioni in dataset non-lineari

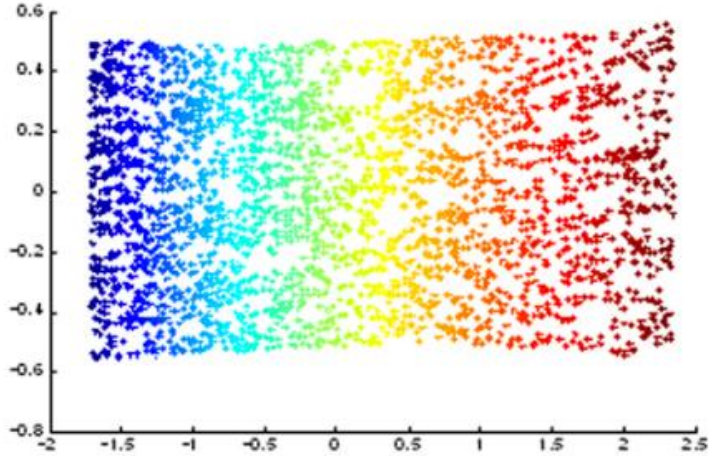
ISOMAP

L'**Isomap** è un metodo di dimension reduction che preserva la geometria *locale* di un dataset

- Per ogni punto si calcola un k-neighborhood. **Nel vicinato la distanza verrà preservata.**
- Per ogni vicinato si costruisce un grafo. la distanza tra due punti equivale al percorso più breve nel grafo (**stima della distanza geodetica**)
- Si rappresenta il grafo connesso in un nuovo sistema di riferimento gerarchico (**Multi-dimensional Scaling**)

Isomap non preserva globalmente le distanze





- GRAZIE

Via Dante, 6, 21052 Busto Arsizio VA
Tel.: +39.0331.357.400
Fax: +39.0331.622.869

Reti