COESIONE ITALIA 21-27 EMILIA-ROMAGNA

Cofinanziato dall'Unione europea

GUARDA AVANTI

Big Data, nuove competenze per nuove professioni

(Progetto rivolto a laureati in tutte le aree disciplinari, co-finanziato dal Fondo Sociale Europeo Plus 2021-2027 Regione Emilia-Romagna)

DATA LAB

Programma della lezione

- Riduzione della dimensionalità
- PCA
- t-SNE

Classical Machine Learning

Task Driven

Data Driven

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

Unsupervised Learning

(Unlabelled Data)

Clustering

(Divide by Similarity)

Eg. Targeted Marketing

Association

(Identify Sequences)

Eg. Customer Recommendation

Dimensionality Reduction

(Wider Dependencies)

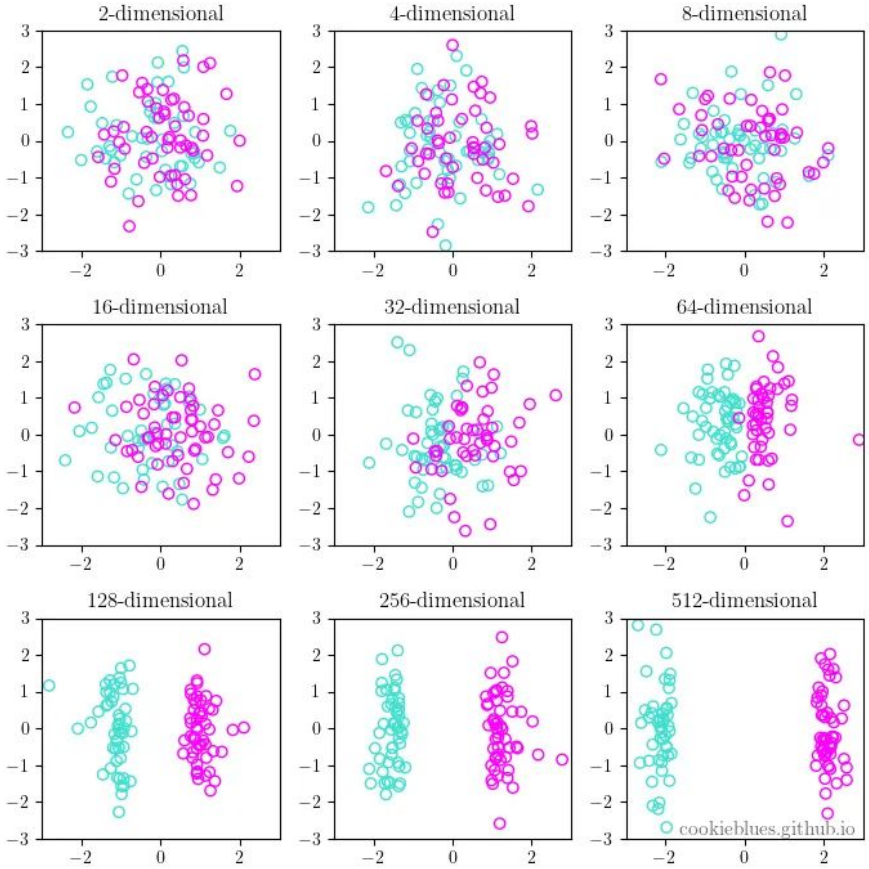
Eg. Big Data Visualization

Obj: Predications & Predictive Models

Pattern/ Structure Recognition

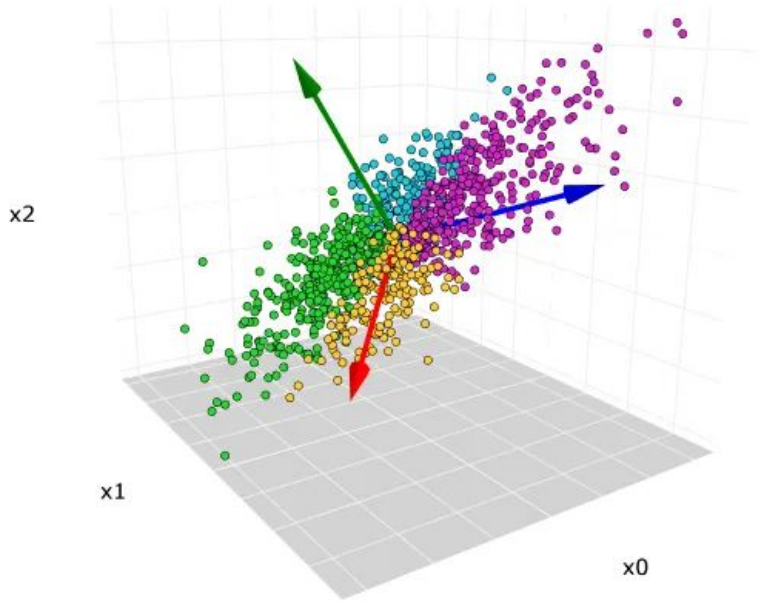
Quando i dati con tante dimensioni cominciano a dare problemi?

- La maledizione della dimensionalità (Curse of Dimensionality) si riferisce a quando hai un dataset con troppe variabili
- Se tu hai più variabili che osservazioni, c’è un alto rischio di overfitting quando costruisci il modello
- Se il dataset contiene troppe variabili, c’è difficoltà nell’individuare i gruppi dalle osservazioni


2-dimensional
4-dimensional
8-dimensional
16-dimensional
32-dimensional
64-dimensional
128-dimensional
256-dimensional
512-dimensional
cookieblues.github.io

Principal Component Analysis

- È una tecnica che trasforma dati ad alta dimensionalità in bassa dimensionalità, cercando di conservare più informazione possibile
- PCA è estremamente utile quando si lavora con dataset che hanno **tante variabili**



Principal Component Analysis

- Si pone due domande:
  - Come possiamo capire quale parte dei nostri dati è più importante?
  - Come possiamo quantificare matematicamente la mole di informazioni?

Principal Component Analysis

- Si pone due domande:
  - Come possiamo capire quale parte dei nostri dati è più importante?
  - Come possiamo quantificare matematicamente la mole di informazioni?

→ La Varianza può!

Cos’è la Varianza?

Misura quanto ogni osservazione differisce dalla media

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Esempio di Varianza

Person	Height (cm)
Alex	145
Ben	160
Chris	185

Esempio di Varianza

Person	Height (cm)
Daniel	172
Elsa	173
Fernandez	171

Cosa abbiamo capito?

- Quando i nostri dati presentano una varianza maggiore, contengono più informazioni
- Per questo motivo, l’obiettivo della PCA è preservare più varianza possibile

Se aggiungessimo una nuova variabile?

Person	Height (cm)	Weight (kg)
Alex	145	68
Ben	160	67
Chris	185	69

Se aggiungessimo una nuova variabile?

Person	Height (cm)	Weight (kg)
Alex	145	68
Ben	160	67
Chris	185	69

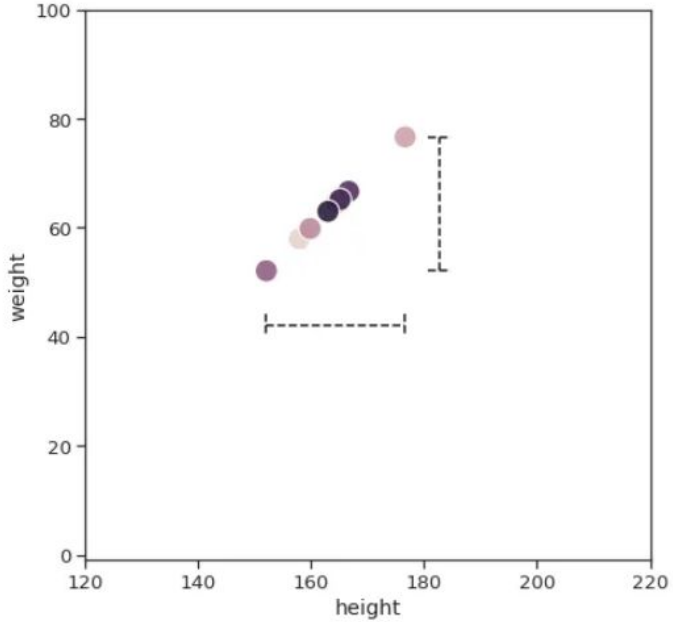
Le differenze di peso non aiutano a differenziare i nostri amici!

Se aggiungessimo una nuova variabile?

Person	Height (cm)	Weight (kg)
Alex	145	68
Ben	160	67
Chris	185	69

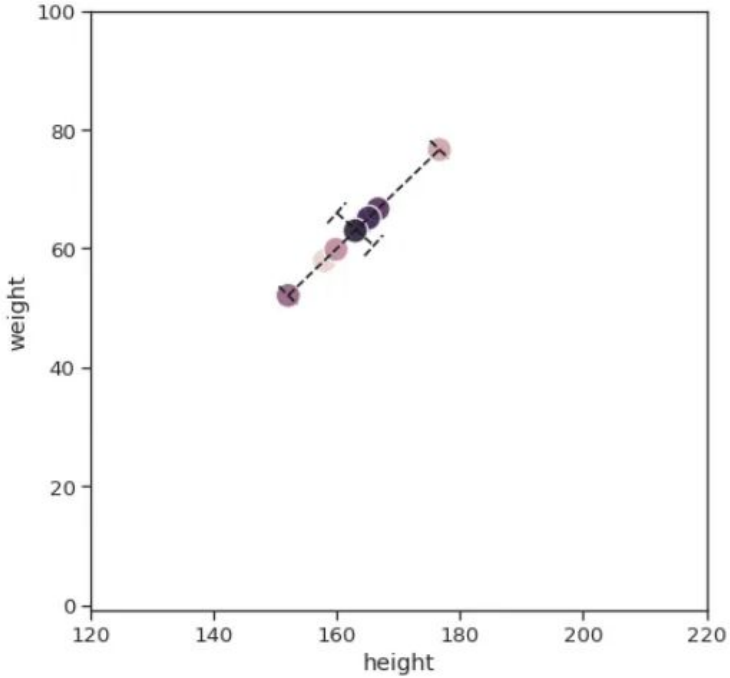
- Le differenze di peso non aiutano a differenziare i nostri amici!
- Con questo ragionamento, stiamo riducendo i nostri dati da 2 a una dimensione

Quale variabile è la più informativa?



Feature	Variance
Height	1.11
Weight	1.11
TOTAL	**2.22**

La quantità massima di varianza si trova …

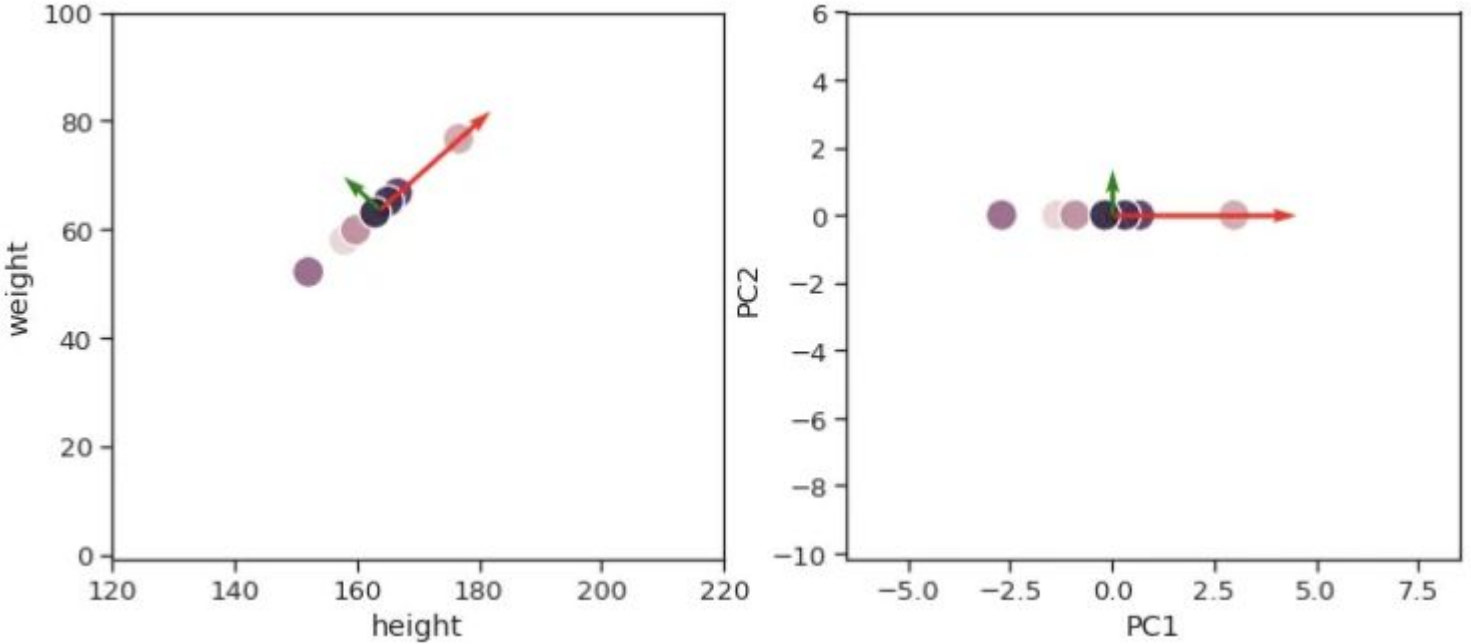


Feature	Variance
Height	1.11
Weight	1.11
TOTAL	**2.22**

Cosa ha fatto?

- Per rappresentare queste due linee, PCA ha combinato altezza e peso per creare due variabili completamente nuove!
- Queste due nuove variabili si chiamano
  - Prima componente principale
  - Seconda componente principale

Prima e seconda componente principale



Come seleziono le componenti principali?

Feature	Variance	Feature	Variance
Height	1.11	PC1	2.22
Weight	1.11	PC2	0.00
TOTAL	**2.22**	**TOTAL**	**2.22**

Come seleziono le componenti principali?

Feature	Variance	Feature	Variance
Height	1.11	PC1	2.22
Weight	1.11	PC2	0.00
TOTAL	**2.22**	**TOTAL**	**2.22**

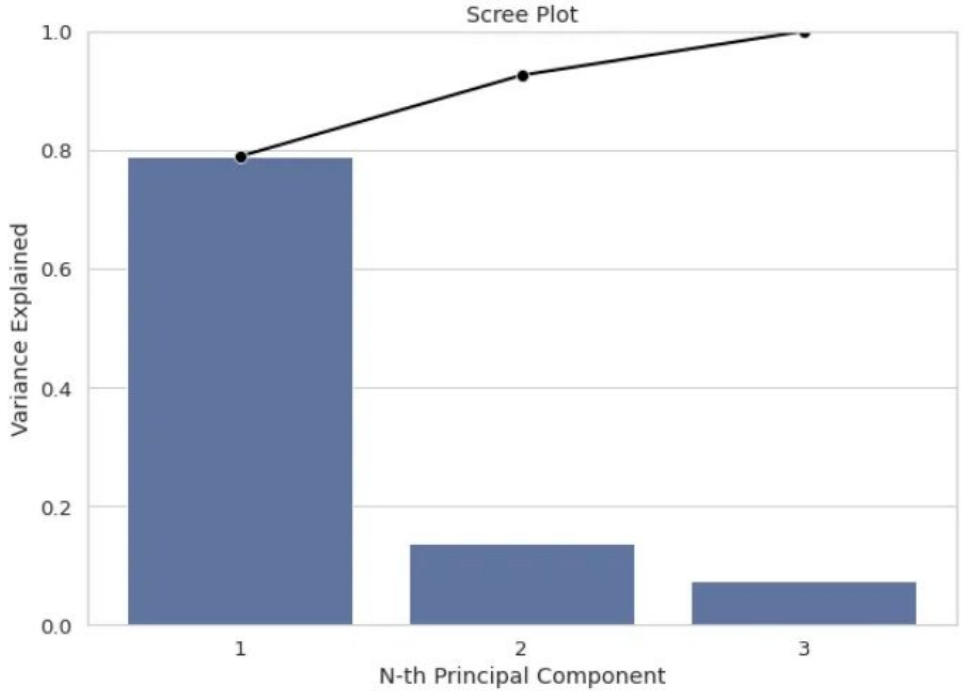
Seleziono PC1 perché è in grado di catturare la varianza totale di altezza e peso combinati

Cosa succede nella realtà?

- In un dataset con tante colonne, è molto difficile ottenere una componente principale che catturi il 100% delle varianze
- Bisognerà selezionare n componenti principali dove n < N
  - N = dimensione dei nostri dati originali
- Requisiti:
  - Bisogna scegliere il minor numero possibile di componenti
  - Catturare una variabilità intorno al 70-80%

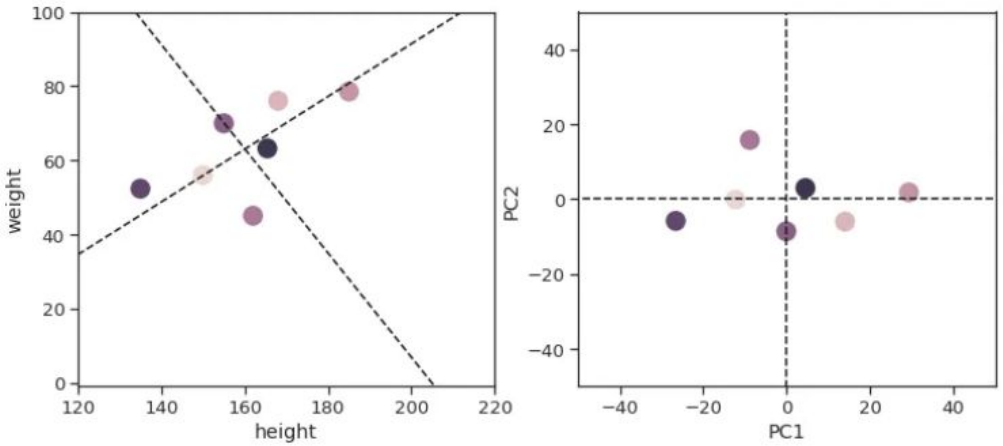
Come si selezionano le componenti principali?

- Tramite lo scree plot
- Guardando:
  - Proporzione di varianza spiegata da ogni componente principale
  - Somma cumulata della varianza fino al componente n-esima



Quindi cosa fa la PCA?

- Quando trasformiamo i nostri dati originali da 2 dimensioni a 2 dimensioni, tutto rimane uguale tranne l'**ORIENTAMENTO**!
- Abbiamo semplicemente ruotato i nostri dati in modo che la varianza sia massima in PC1

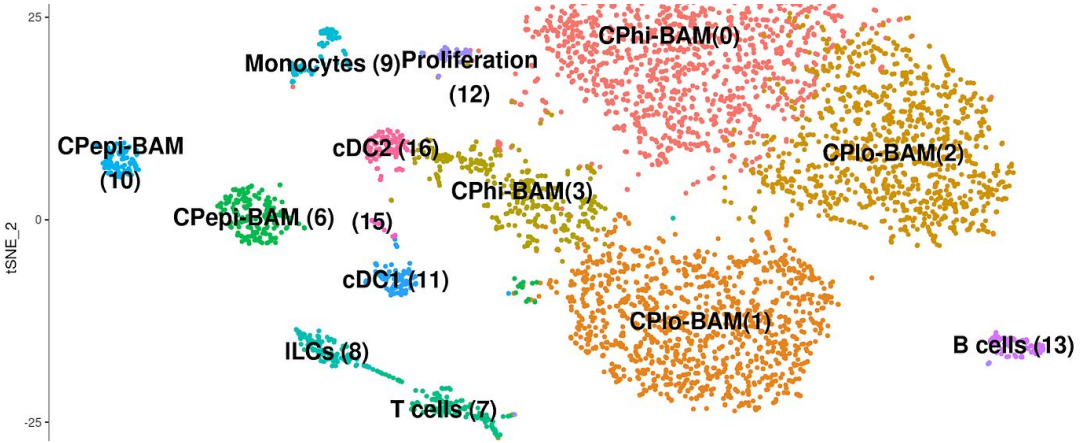


Vantaggi della PCA

- Riduzione della dimensionalità
- Aiuta a visualizzare dati con tante colonne
- È efficiente computazionalmente
- Ma ha anche svantaggi
  - Sensibili agli outlier
  - Non sempre la varianza è sinonimo di informazione
  - La PCA assume relazioni lineari tra le variabili e potrebbe non funzionare bene con dati non lineari

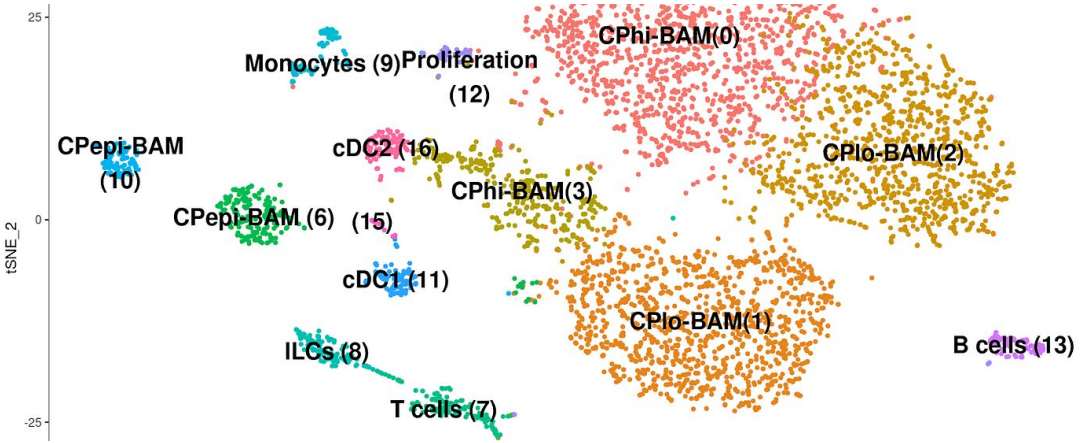
t-SNE

- L’acronimo di t-SNE è t-distributed stochastic neighbor embedding
  - Embedding
  - Neighbor - punto che è vicino al punto di interesse
  - Stocastico per l’uso della causalità nel processo iterativo
  - T-distributed è la distribuzione utilizzata dall’algoritmo per calcolare gli score di similarità nei dati di dimensione inferiore



Idee della t-SNE

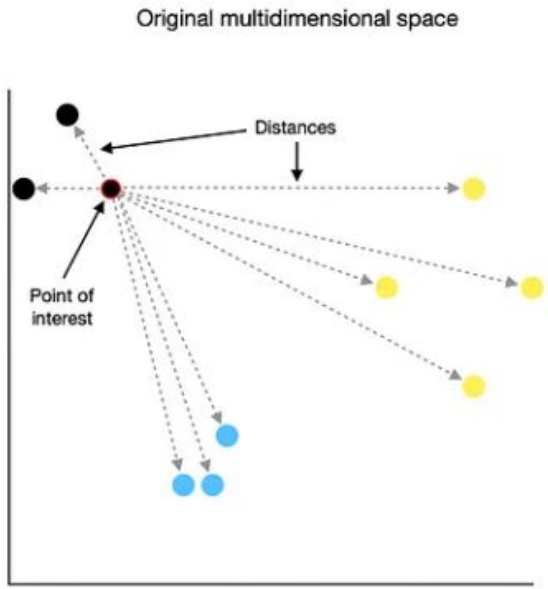
- Trova un modo di proiettare i dati in uno spazio a bassa dimensione in modo che i **cluster nello spazio multidimensionale vengano preservati**



Fase 1

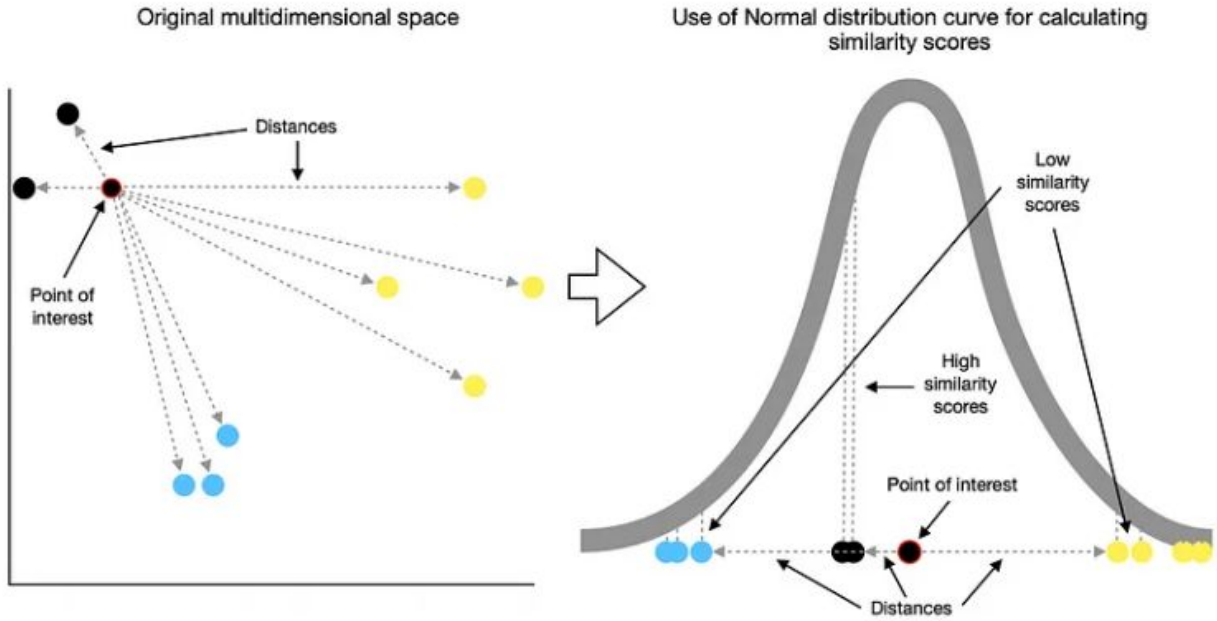
- t-SNE inizialmente determina la similarità tra i punti in base alla distanza tra loro
  - Punti vicini sono considerati “simili”
  - Punti lontani sono considerati “dissimili”

Fase 1



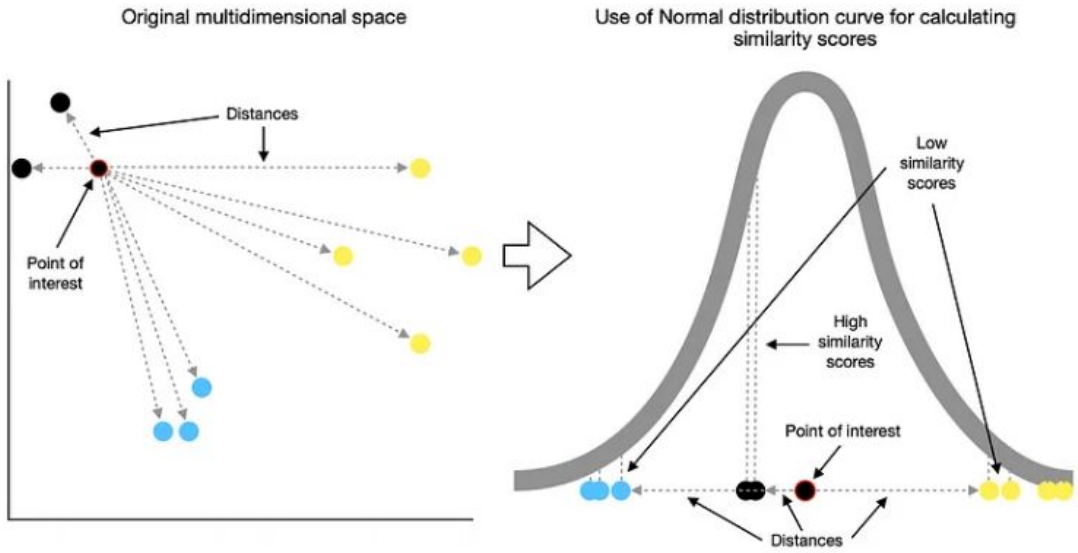
Calcola le distanze tra il punto di interesse e gli altri punti

Fase 1



Dopo averle calcolate, li posiziona su una curva normale

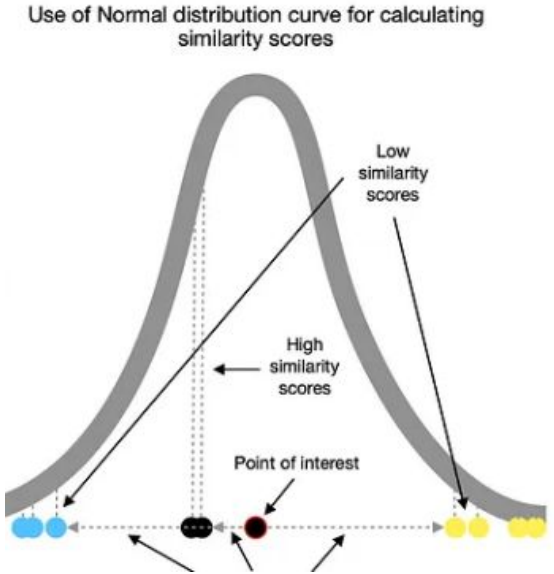
Fase 1



Dopo traccia una linea tratteggiata dal punto alla curva. Questa linea tratteggiata è la similarità calcolata tra il punto di interesse e il punto preso in considerazione

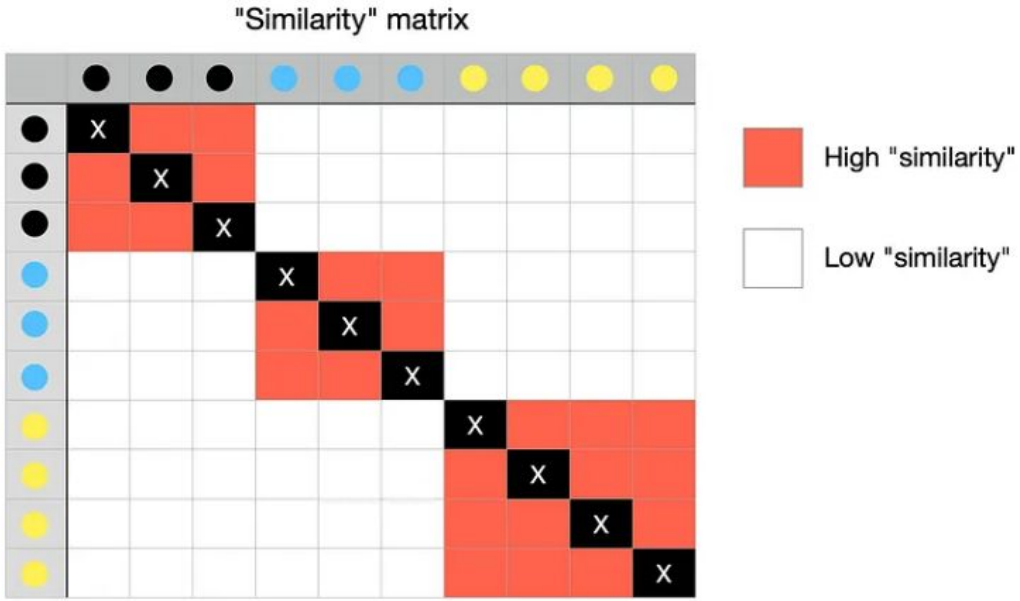
Perché usare la distribuzione normale?

- I punti distanti avranno un valore di similarità molto basso
- I punti vicini avranno un valore di similarità molto alto



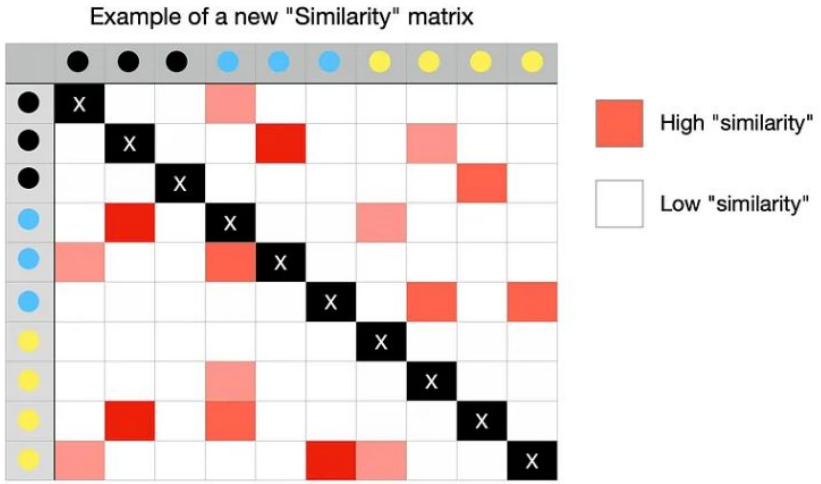
Fase 1

- Il risultato di questi calcoli è una matrice che contiene tutti gli score di similarità tra ogni coppia di punti nello **spazio multidimensionale originale**

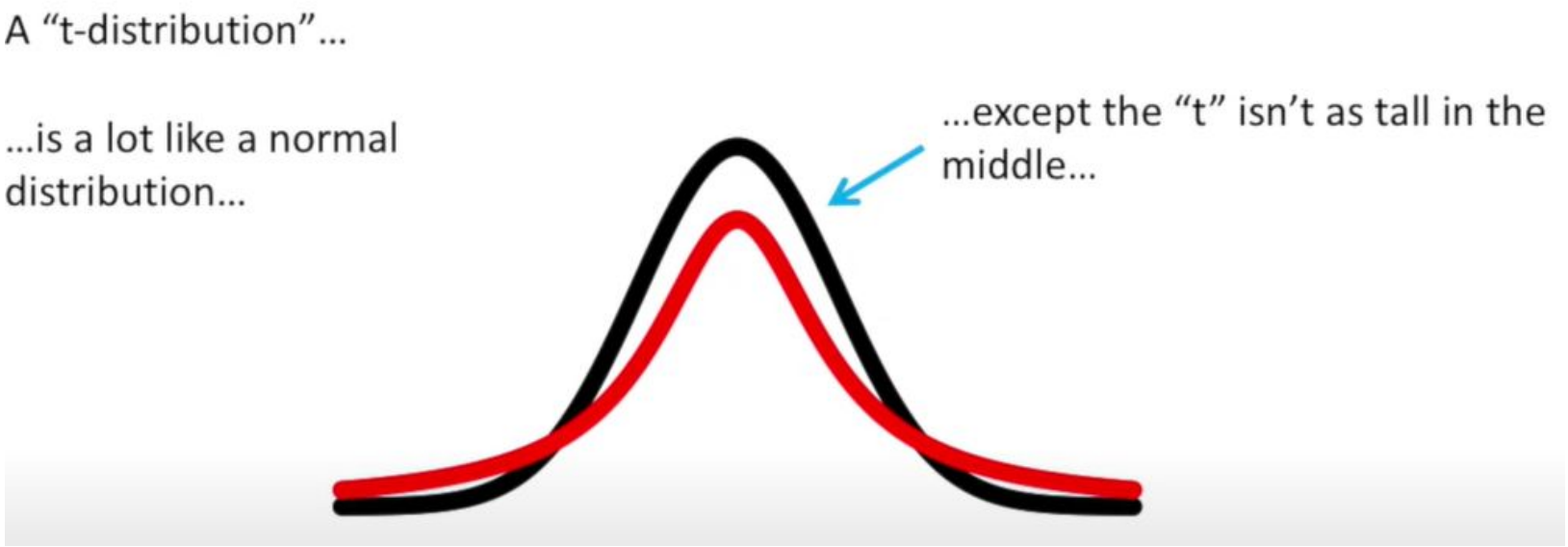


Fase 2

- t-SNE mappa casualmente tutti i punti su uno spazio a dimensione inferiore
- Calcola gli score di similarità tra i punti come nel processo di prima
  - Calcola le distanze tra il punto di interesse e gli altri punti
  - Posiziona ogni coppia di punti sulla curva della distribuzione t-student
  - Traccia una linea tratteggiata dal punto alla curva
- Ottengo la nuova matrice di similarità



Distribuzione Normale Vs Distribuzione t-student

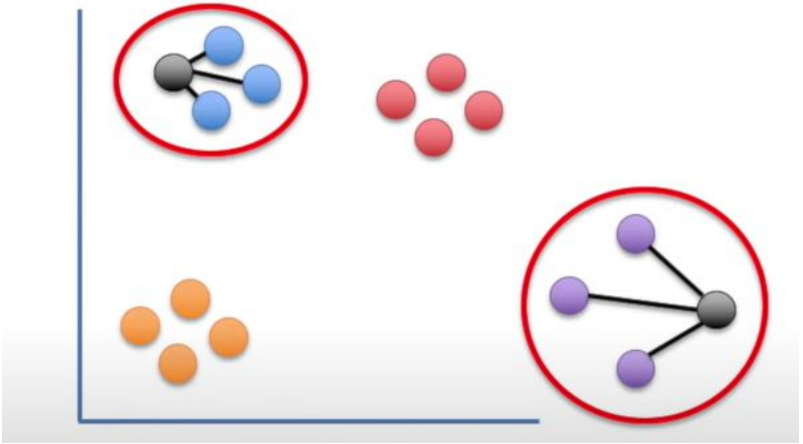


Fase 3

- L’obiettivo è far sì che la nuova matrice di similarità assomigli alla matrice di similarità calcolata sui dati originali utilizzando un processo iterativo
  - Ad ogni iterazione, ogni punto si muoverà verso i punti a cui era più vicino nei dati originali e starà lontano da quelli a cui era distante
- Dopo un certo numero di iterazioni, la nuova matrice di similarità, calcolata in Fase 2, assomiglierà sempre di più a quella originale
- Il processo si ferma quando
  - Il numero massimo di iterazioni è raggiunto
  - Non è possibile apportare ulteriori miglioramenti

Parametri della t-SNE

- La densità prevista intorno a ciascun punto si chiama perplexity
- Su sklearn perplexity è il numero di punti vicini al punto di interesse
- Di default è 30 su sklearn
- Più grande è il dataset, maggiore sarà il parametro
- Perplexity è definito dall'utente



Vantaggi della t-SNE

- Preservazione delle strutture locali
- Visualizzazione intuitiva
- Flessibile con dati non lineari
- Ma …
- È difficile da interpretare
- Alto costo computazionale
- Non deterministico
- Sensibili ai parametri

Resources

- PCA tutorial
- t-SNE tutorial