Apprendimento Non Supervisionato: Riduzione della Dimensionalità

Docente: Tommaso Muraca

Riduzione della Dimensionalità

La **riduzione della dimensionalità** assomiglia molto alla **compressione**, si tratta di **cercare di ridurre la complessità dei dati** mantenendo quanto più possibile la **struttura rilevante**. Se prendiamo una semplice immagine di 128 x 128 x 3 pixel (lunghezza x larghezza x valore RGB), si tratta di 49.152 dimensioni di dati. Se riusciamo a ridurre la dimensionalità dello spazio in cui vivono queste immagini senza distruggere troppo il contenuto significativo delle immagini, allora abbiamo fatto un buon lavoro nella riduzione della dimensionalità.

Daremo uno sguardo a **due tecniche** comuni nella pratica: **l'analisi delle componenti principali** e la **scomposizione dei valori singolari** .

- Analisi delle componenti principali (PCA)

Innanzitutto, un piccolo ripasso di algebra lineare: parliamo di spazi e basi .

Abbiamo familiarità con il piano delle coordinate con origine O(0,0) e vettori base i(1,0) e j(0,1).

Si scopre che possiamo scegliere una base completamente diversa e avere comunque tutti i conti risolti. A esempio, possiamo mantenere O come origine e scegliere la base dei vettori i'=(2,1) e j'=(1,2).

Vedremo che il punto etichettato (2,2) nel sistema di coordinate i', j' è etichettato (6, 6) nel sistema i, j.

Riduzione della Dimensionalità: PCA

Ciò significa che possiamo **cambiare la base di uno spazio**. Ora immaginiamo uno spazio con dimensioni molto più elevate. A esempio, **dimensioni 50K**, possiamo selezionare una **base per quello spazio**, quindi selezionare solo i **200 vettori più significativi di quella base**.

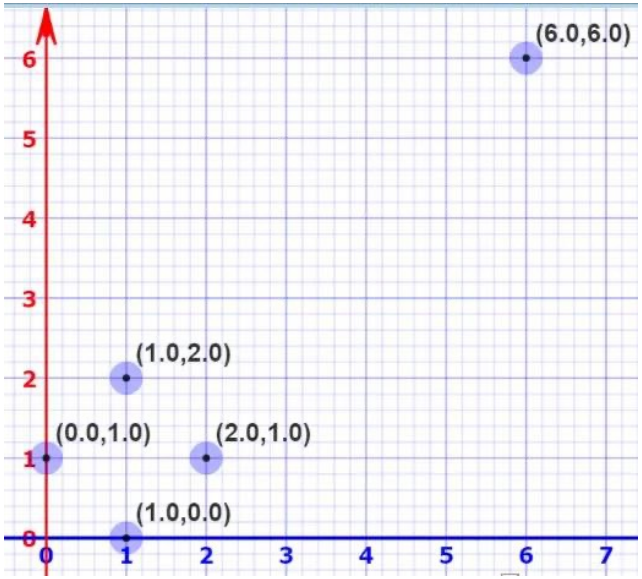
Questi vettori di base sono **chiamati componenti principali** e il sottoinsieme selezionato costituisce un nuovo spazio che ha **dimensioni inferiori rispetto allo spazio originale** ma mantiene la **massima complessità dei dati possibile.**

Per **selezionare** le **componenti principali più significative**, esaminiamo quanta **varianza dei dati catturano** e le **ordiniamo in base a tale metrica**.

Un altro modo di pensare a questo è che **PCA rimappa lo spazio in cui esistono i nostri dati** per renderli più comprimibili. La dimensione trasformata è più piccola della dimensione originale.

Utilizzando solo le **prime dimensioni dello spazio rimappato**, possiamo iniziare a comprendere l'organizzazione del set di dati.

Questa è la **promessa della riduzione della dimensionalità**: ridurre la complessità (dimensionalità in questo caso) mantenendo la struttura (varianza).



Riduzione della Dimensionalità: SVD

- Scomposizione in valori singolari (SVD)

Rappresentiamo i nostri dati come una **grande matrice A = m x n** .

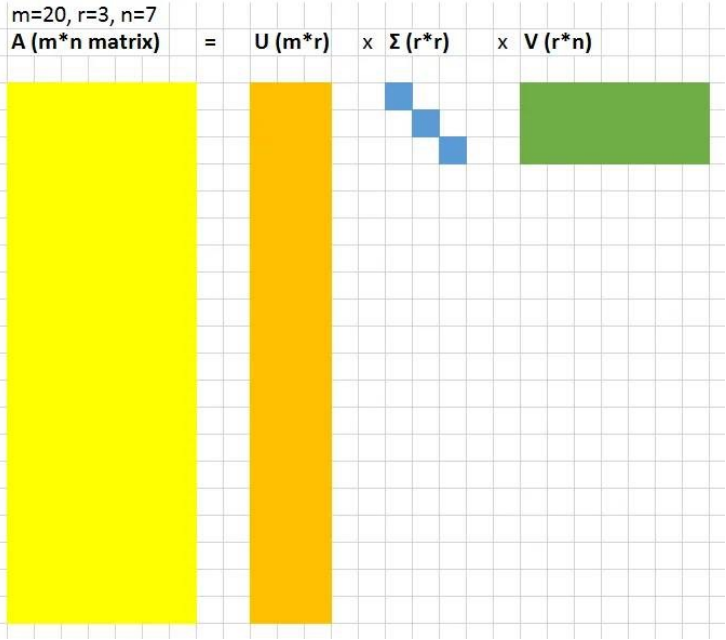
SVD è un calcolo che ci consente di **scomporre quella grande matrice in un prodotto di 3 matrici più piccole** (U= m x r , matrice diagonale Σ= r x r , e V= r x n dove r è un numero piccolo).

A destra un'illustrazione più visiva di quel prodotto per iniziare -->

I valori nella matrice diagonale r*r Σ sono detti **valori singolari.**

La cosa bella è che questi valori singolari **possono essere usati per comprimere la matrice originale.**

Se si elimina il 20% più piccolo dei valori singolari e le colonne associate nelle matrici U e V, si risparmia un bel po' di spazio e si ottiene comunque una rappresentazione decente della matrice sottostante.

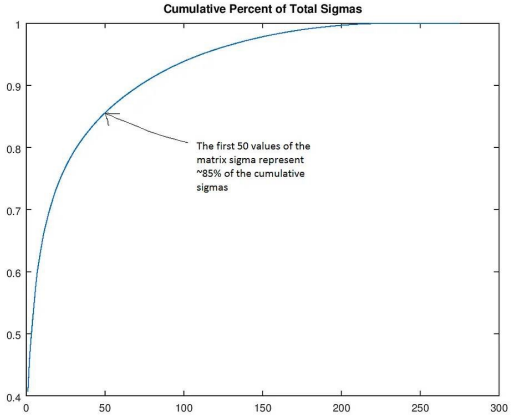


Riduzione della Dimensionalità: SVD

Per esaminare cosa significa più precisamente, **lavoriamo con questa immagine di un cane →**

Innanzitutto, mostriamo che **se classifichiamo i valori singolari** (i valori della matrice Σ) **in base alla grandezza**, i **primi 50 valori singolari** contengono **l'85% della grandezza dell'intera matrice Σ.**

Possiamo sfruttare questo fatto per **scartare i successivi 250 valori di sigma** (cioè impostarli su 0) e mantenere semplicemente una **versione di "rank 50"** dell'immagine del cane.



Nell'immagine della slide successiva creiamo un **cane di grado 200, 100, 50, 30, 20, 10 e 3**. Ovviamente il quadro è più piccolo, ma concordiamo sul fatto che il **cane di grado 30 è comunque buono.** Ora vediamo quanta compressione otteniamo con questo cane.

La matrice dell'immagine originale è 305*275 = 83.875 valori.

Riduzione della Dimensionalità: SVD

Il cane di rango 30 è 305*30+30+30*275=17.430: quasi **5 volte meno valori** con una **perdita minima nella qualità dell'immagine.**

Il motivo del calcolo sopra è che **scartiamo anche le parti della matrice U e V che vengono moltiplicate per zeri** quando viene eseguita l'operazione U Σ' V (dove Σ' è la versione modificata di Σ che ha solo i primi 30 valori dentro).

L'apprendimento non supervisionato viene spesso utilizzato per **preelaborare i dati**, di solito ciò significa **comprimerlo in qualche modo preservandone il significato**, come con PCA o SVD, prima di **inviarlo a una rete neurale profonda** o a un altro **algoritmo di apprendimento supervisionato.**

Ora come al solito andiamo alla pratica!