Principal Component Analysis

Principal Component Analysis

Immaginiamo di avere una matrice A ∈ Rn×m contenente dei dati. In particolare di avere m

campioni e per ogni campione misuriamo n variabili: esempio altezza e peso o altezza ed età.

* Il primo passo consiste nel calcolare la media campionaria lungo ogni riga della matrice A = ∑ Aij/m.
* Successivamente si sottrae la media da tutte le entrate di una riga. In questo modo ogni riga di una nuova matrice centrata A ha media zero: A = A − A’(1, . . . , 1)T .
* Le colonne di A saranno fatte da n punti e avremo m colonne. A causa dell’operazione di centramento, la somma degli m vettori colonna è zero. Di conseguenza la colonna media coincide con il vettore nullo.
* Spesso quello che si osserva è che questi insiemi di dati sono concentrati intorno a un sottospazio di bassa dimensione di Rn: una retta, un piano ad esempio.
* Come si trova con gli strumenti dell’algebra lineare la retta per (0, 0) più vicina alla nuvola di punti?
* Questa retta è semplicemente lungo la direzione del primo valore singolare u1 di A. Questo è il punto chiave di tutta la PCA (Principal Component Analysis).

Vediamo il problema da un punto di vista statistico.

Supponiamo ad esempio che le colonne di A corrispondano all’età (prima riga) e all’altezza

(seconda riga) di un insieme di 6 bambini, misurate a partire dall’età e altezza media

(misurano quanto si discostano dalla quantità media).

A = [3 −4 7 1 −4 −3; 7 −6 8 −1 −1 7]

Stiamo ora cercando una retta che rimanga più vicino possibile all’insieme dei dati. Facciamo ricorso alla varianza e alla covarianza:

* La varianza è rappresentata dalla somma delle distanze dalla media lungo ogni riga della matrice A al quadrato: 1/(m-1) Σ x2, dove xi sono i valori sulle righe.
* Per ogni riga la varianza è rappresentata dalle entrate diagonali della matrice AAT, la covarianza dai termini fuori dalla diagonale di AAT.
* Alta covarianza significa che ad un aumento di altezza corrisponde un aumento di età. Covarianza negativa significa che una variabile cresce mentre l’altra aumenta.

Nell’ esempio la matrice simmetrica AAT è una 2x2 . La matrice della covarianza è definita come S = AAT/(m−1) e vale S = [20 25; 25 40].

I due autovettori ortogonali di S sono u1 e u2. Questi sono anche i vettori singolari, ovvero le componenti principali, di A.

Il Teorema di Eckart-Young ci dice che il vettore u1 è diretto lungo la retta che è più vicina ai dati. Il secondo valore singolare u2 è ortogonale alla retta che minimizza la distanza dai dati.

In generale, la PCA è descritta dalla matrice simmetrica S = AAT/(n − 1) ma dati i valori di A, il calcolo della decomposizione SVD fornisce valori più precisi e in maniera più rapida.

Nell’esempio, S ha autovalori vicino a 57 e 3. La prima matrice di rango 1 `e √57u1v1 che è molto più grande della seconda di rango unitario √3u2v2. L’autovettore principale u1 ≈ (0.6, 0.8) ci dice che la retta più vicina ai dati ha pendenza circa 8/6.

Analisi della PCA: punto di vista geometrico

Cerchiamo di capire in che senso la retta che ha la direzione di u1 è la retta più vicina ai

valori centrati indicati dalla matrice A?

Questa retta cercata risolve un problema ai minimi quadrati perpendicolari, chiamata regressione ortogonale. È diversa dai minimi quadrati dove si cerca il minimo di ∥Ax − b∥2 misurando le distanze in alto e in basso dalla retta di miglior approssimazione.

In questo problema invece minimizziamo le distanze perpendicolari alle direzioni principali.

Si cerca la retta per la quale la somma delle distanze al quadrato dei punti dalla retta u1

sia minima.

Separando ogni colonna aj di A nelle sue componenti lungo u1 e u2 si ottiene:

∑ ||aj||2 = ∑ |aju1|2 + ∑ |aju2|2

si vede che la somma a sinistra è fissata, mentre il primo termine a destra vale u1’ajajtu1.

Perciò quando massimizziamo la prima ∑, prendendo l’autovettore u1 di AAT, minimizziamo la seconda somma.

Questa seconda somma definisce le distanze al quadrato dall’insieme di punti dalla retta

per u1 ed è la più piccola possibile.

Modelli matematici