7.2 Analisi delle componenti principali (PCA): codifica ottimale dato il sottospazio

Sia $X \in \mathbb{R}^{m\times d}$ una matrice dati e si supponga di aver fissato (in qualche modo) un sottospazio $\mathcal{S}$ di $\mathbb{R}^d$. Sia $K \leq d$ la dimensione di tale sottospazio e siano $c_1, \ldots, c_K \in \mathbb{R}^d$ vettori che compongono una base di $\mathcal{S}$. Chiamiamo C la matrice $d \times K$ le cui colonne sono i vettori $c_1, \ldots, c_K$.

Il sottospazio $\mathcal{S}$ è tanto più rappresentativo dei dati quanto più la quantità

$$g(C) = \frac{1}{m}\sum_{i=1}^{m}\left\|CC^\top x^{(i)} - x^{(i)}\right\|^2$$

è piccola. Infatti l'operatore $CC^\top$ è proprio l'operatore di proiezione sul sottospazio $\mathcal{S}$: è un operatore lineare, ed è idempotente perché

$$(CC^\top)(CC^\top) = C(C^\top C)C^\top = CC^\top,$$

dove la seconda uguaglianza segue dal fatto che $C^\top C$ è una matrice identità, visto che i vettori $c_1, \ldots, c_K$ formano una base ortonormale di $\mathcal{S}$. Infine, qualunque elemento $u \in \mathcal{S}$ è della forma $u = Cv$ per qualche $v \in \mathbb{R}^K$, e quindi

$$CC^\top u = CC^\top(Cv) = Cv = u,$$

quindi la proiezione lascia invariati tutti gli elementi di $\mathcal{S}$.

7.3 Analisi delle componenti principali (PCA): prima componente principale

Qual è una matrice a colonne ortogonali $C \in \mathbb{R}^{d\times K}$ (e quindi un sottospazio $\mathcal{S}$) che minimizza la distorsione $g(C)$? Dimostriamo qui che una scelta ottimale della prima colonna c_1 della matrice C è data dall'autovettore principale della ***matrice di covarianza dei dati*** $\Sigma = (1/m)X^\top X$.

Poiché $CC^\top$ è un operatore idempotente,

$$\begin{aligned}\left\|CC^\top x^{(i)} - x^{(i)}\right\|^2 &= x^{(i)\top} CC^\top CC^\top x^{(i)} - 2x^{(i)\top} CC^\top x^{(i)} + x^{(i)\top} x^{(i)} \\ &= -x^{(i)\top} CC^\top x^{(i)} + x^{(i)\top} x^{(i)} \\ &= -\left\|C^\top x^{(i)}\right\|^2 + \left\|x^{(i)}\right\|^2.\end{aligned}$$

Visto che $\left\|x^{(i)}\right\|^2$ non dipende da C,

$$\begin{aligned}\operatorname*{argmin}_C g(C) &= \operatorname*{argmin}_C \left(-\frac{1}{m}\sum_{i=1}^{m}\left\|C^\top x^{(i)}\right\|^2\right) \\ &= \operatorname*{argmax}_C \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{K}\langle c_j, x^{(i)}\rangle^2\end{aligned}$$

$$= \underset{C}{\operatorname{argmax}} \frac{1}{m} \sum_{j=1}^{K} \sum_{i=1}^{m} \langle c_j, x^{(i)} \rangle^2,$$

dove si è usato il fatto che la j-esima componente di $C^\top x^{(i)}$ è $\langle c_j, x^{(i)} \rangle$.

Una delle colonne di C, ad esempio c_1, può essere scelta indipendentemente dalle altre, e quindi può essere scelta in modo da massimizzare $(1/m) \sum_{i=1}^{m} \langle c_1, x^{(i)} \rangle^2$. Infine osserviamo che

$$\frac{1}{m} \sum_{i=1}^{m} \langle c_1, x^{(i)} \rangle^2 = c_1^\top \left(\frac{1}{m} X^\top X \right) c_1,$$

dove si è usato il fatto che

$$X c_1 = \begin{pmatrix} \langle c_1, x^{(1)} \rangle \\ \langle c_1, x^{(2)} \rangle \\ \cdots \\ \langle c_1, x^{(m)} \rangle \end{pmatrix}.$$

Quindi la scelta ottima di c_1 consiste nell'autovettore principale della matrice di covarianza Σ, in quanto il massimizzante della funzione $u \mapsto u^\top S u$ è l'autovettore principale della matrice S, qualunque sia la matrice semidefinita positiva S.

Si noti che una volta fissata la colonna c_1, le altre colonne non possono più essere scelte liberamente, in quanto devono soddisfare la condizione di ortogonalità: c_2 dovrà essere ortogonale a c_1, e in generale c_j dovrà essere ortogonale al sottospazio generato da $c_1, \ldots, c_{j-1}$.