Lezione 20: Analisi delle componenti principali.

Una forma comune di UNSUPERVISED LEARNING è la riduzione dimensionale in cui si impara una mappa tra lo spazio visibile ad alta dimensione, $\vec{x} \in \mathbb{R}^D$, ad uno spazio di dimensione più piccola (latent space), $\vec{z} \in \mathbb{R}^L$.
La forma più semplice di riduzione dimensionale è l'analisi delle componenti principali (PRINCIPAL COMPONENT ANALYSIS). L'idea è di trovare una proiezione lineare e ortogonale dello spazio $\mathbb{R}^D$ su uno spazio $\mathbb{R}^L$ tale che quest'ultimo sia una "buona approssimazione" del primo. Per "buona approssimazione" si intende che se proiettiamo (ENCODE) $\vec{x}$ per ottenere $\vec{z} = W^T \vec{x}$ e poi applichiamo l'operazione inversa (DECODE) per ottenere $\hat{\vec{x}} = W\vec{z}$, allora $\vec{x}$ e $\hat{\vec{x}}$ sono vicini (nel senso di distanza "euclidea" l_2).
Definiamo il RECONSTRUCTION ERROR o DISTORTION come:

$$\mathcal{L}(W) \stackrel{def}{=} \frac{1}{N_D} \sum_{n=1}^{N_D} \| x_n - \text{decode}(\text{encode}(x_n; W); W) \|^2_{l_2}$$

dove gli stadi "ENCODE" e "DECODE" sono mappe lineari.
Vogliamo dunque definire una procedura che minimizzi $\mathcal{L}(W)$. Questa procedura è tale se $\hat{W} = U_L$ dove U_L contiene gli L autovettori con gli autovalori maggiori [e $\hat{W}$ è la trasformazione che minimizza $\mathcal{L}(W)$] della matrice di covarianza empirica

$$\hat{\Sigma} = \frac{1}{N_D} \sum_{n=1}^{N_D} (x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N} X_c^T X_c$$

Esempio: i punti nello spazio proiettati su una retta

Supponiamo di avere un data set (unlabeled)

$$\mathcal{D} = \{\vec{x}_n\} \quad n = 1, \dots, N_{\mathcal{D}} \quad \text{con} \quad \vec{x}_n \in \mathbb{R}^D$$

Questo data set può essere rappresentato da una matrice $X_{N \times D}$

Definiamo $\bar{x} = \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \vec{x}_n = \vec{0}$ che può essere ottenuto centrando i dati

Il nostro scopo è di approssimare $\vec{x}_n$ con uno $\vec{z}_n \in \mathbb{R}^L$. Assumiamo che ciascun $\vec{x}_n$ possa essere "spiegato" in termini di una combinazione di funzioni della base $\vec{w}_1, \dots, \vec{w}_L$ dove ciascun $\vec{w}_k \in \mathbb{R}^L$ e dove i pesi sono dati da $\vec{z}_n \in \mathbb{R}^L$ cioè:

$$\vec{x}_n \simeq \sum_{k=1}^{L} z_{nk} \vec{w}_k$$

Il vettore $\vec{z}_n$ è la rappresentazione (low-dimensional) di $\vec{x}_n$ ed è noto come "LATENT VECTOR" (per via del fatto che consiste in valori "nascosti" non osservati nei dati). L'errore può dunque essere misurato come:

$$\mathcal{L}(W, Z) = \frac{1}{N} \| X - Z W^T \|^2 = \frac{1}{N} \| X^T - W Z^T \|^2 = \frac{1}{N} \sum_{n=1}^{N_{\mathcal{D}}} \| \vec{x}_n - W \vec{z}_n \|^2$$

Vogliamo minimizzare quest'espressione col vincolo che W sia una matrice ortogonale.

Iniziamo stimando la miglior soluzione monodimensionale $\vec{w}_1 \in \mathbb{R}^D$

Denotiamo con $\tilde{z}_1 = \{z_{11}, z_{21}, \dots z_{N1}\} \in \mathbb{R}^{N_{\mathcal{D}}}$ i coefficienti per ciascuno dei vettori $\vec{x}_n$ associati al primo vettore della base. Allora

$$\mathcal{L}(\vec{w}_1, \tilde{\vec{z}}_1) = \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \| \vec{x}_n - z_{n1} \vec{w}_1 \|^2 = \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} (\vec{x}_n - z_{n1} \vec{w}_1)^T (\vec{x}_n - z_{n1} \vec{w}_1)$$

$$= \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \left[\vec{x}_n^T \vec{x}_n - 2 z_{n1} \vec{w}_1^T \vec{x}_n + z_{n1}^2 \underbrace{\vec{w}_1^T \vec{w}_1} \right]$$

$$= \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \left[\vec{x}_n^T \vec{x}_n - 2 z_{n1} \vec{w}_1^T \vec{x}_n + z_{n1}^2 \right] \quad 1 \text{ in virtù dell'ortonormalità}$$

Consideriamo ora le derivate rispetto a z_{n1}

$$\frac{\partial}{\partial z_{n1}} \mathcal{L}(\vec{w}_1, \tilde{z}_1) = \frac{1}{N_D}\left[-2\vec{w}_1^T \vec{x}_n + 2 z_{n1}\right] = 0 \iff z_{n1} = \vec{w}_1^T \vec{x}_n$$

Quindi la proiezione ottimale si ottiene proiettando ortogonalmente i dati su $\vec{w}_1$. Sostituendo questa soluzione, si ottiene:

$$\mathcal{L}(\vec{w}_1) = \mathcal{L}(\vec{w}_1, \tilde{z}_1^*(\vec{w}_1)) = \frac{1}{N_D}\sum_{n=1}^{N_D}\left[\vec{x}_n^T \vec{x}_n - z_{n1}^2\right] = \text{const} - \frac{1}{N_D}\sum_{n=1}^{N_D} z_{n1}^2$$

Risolvendo per $\vec{w}_1$

$$\mathcal{L}(\vec{w}_1) = -\frac{1}{N_D}\sum_{n=1}^{N_D} z_{n1}^2 = -\frac{1}{N_D}\sum_{n=1}^{N_D} \vec{w}_1^T \vec{x}_n \vec{x}_n^T \vec{w}_1 = -w_1^T \hat{\Sigma} \vec{w}_1$$

dove $\hat{\Sigma}$ è la matrice di covarianza empirica. Il problema può essere ottimizzato facendo tendere $\|\vec{w}_1\| \to \infty$, quindi imponiamo il constraint $\|\vec{w}_1\| = 1$ e ottimizziamo

$$\tilde{\mathcal{L}}(\vec{w}_1) = w_1^T \hat{\Sigma} \vec{w}_1 - \lambda_1(\vec{w}_1^T \vec{w}_1 - 1)$$

dove λ_1 è un moltiplicatore di Lagrange. Il minimo si trova in

$$\frac{\partial}{\partial \vec{w}_1}\tilde{\mathcal{L}}(\vec{w}_1) = 0 \iff 2\hat{\Sigma}\vec{w}_1 - 2\lambda_1 \vec{w}_1 = 0 \iff \hat{\Sigma}\vec{w}_1 = \lambda_1 \vec{w}_1$$

Quindi, la direzione ottimale su cui proiettare i dati è un autovettore della matrice di covarianza. Moltiplicando a sinistra per $\vec{w}_1^T$ si ottiene (e usando $\vec{w}_1^T \vec{w}_1 = 1$)

$$\vec{w}_1^T \hat{\Sigma} \vec{w}_1 = \lambda_1$$

Poiché vogliamo minimizzare $\mathcal{L}(\vec{w}_1)$, prendiamo la quantità che massimizza $\vec{w}_1^T \hat{\Sigma} \vec{w}_1$, cioè l'autovalore più grande.

<u>Osservazione</u>: poiché i dati sono stati centrati, si ha

$$\mathbb{E}[z_{n1}] = \mathbb{E}[\vec{x}_n^T \vec{w}_1] = \mathbb{E}[\vec{x}_n]^T \vec{w}_1 = 0$$

La varianza dei dati proiettati è:

$$\mathbb{V}[\tilde{z}_1] = \mathbb{E}[\tilde{z}_1^2] - (\mathbb{E}[\tilde{z}_1])^2 = \frac{1}{N_D}\sum_{n=1}^{N_D} z_{n1}^2 - 0 = -\mathcal{L}(\vec{w}_1) + \text{cost.}$$

Vediamo che minimizzare il "RECONSTRUCTION ERROR" è equivalente a massimizzare la varianza dei dati proiettati

$$\underset{\vec{w}_1}{\operatorname{argmin}}\ \mathcal{L}(\vec{w}_1) = \underset{\vec{w}_1}{\operatorname{argmax}}\ \mathbb{V}[\tilde{z}_1(\vec{w}_1)]$$

Questa è la ragione per cui spesso si dice che la PCA trova le direzioni di massima varianza.

Troviamo ora un'altra direzione $\vec{w}_2$ per minimizzare il "RECONSTRUCTION ERROR" che soddisfa la condizione $\vec{w}_1^T\vec{w}_2 = 0$ e $\vec{w}_2^T\vec{w}_2 = 1$. Allora

$$\mathcal{L}(\vec{w}_1, \tilde{z}_1, \vec{w}_2, \tilde{z}_2) = \frac{1}{N_D}\sum_{n=1}^{N_D} \left\| \vec{x}_n - z_{n1}\vec{w}_1 - z_{n2}\vec{w}_2 \right\|^2$$

Si trova che $\frac{\partial \mathcal{L}}{\partial z_2} = 0 \iff z_{n2} = \vec{w}_2^T\vec{x}_n$; sostituendo in $\mathcal{L}$:

$$\mathcal{L}(\vec{w}_2) = \frac{1}{N_D}\sum_{n=1}^{N_D}\left[\vec{x}_n^T\vec{x}_n - \vec{w}_1^T\vec{x}_n\vec{x}_n^T\vec{w}_1 - \vec{w}_2^T\vec{x}_n\vec{x}_n^T\vec{w}_2\right] = \text{const} - \vec{w}_2^T\hat{\Sigma}\vec{w}_2$$

Tralasciamo il termine costante, sostituiamo l'espressione per il $\vec{w}_1$ ottimale e imponiamo il constraint di ortonormalità per ottenere

$$\mathcal{L}(\vec{w}_2) = \vec{w}_2^T\hat{\Sigma}\vec{w}_2 + \lambda_2(\vec{w}_2^T\vec{w}_2 - 1) + \lambda_{12}(\vec{w}_2^T\vec{w}_1 - 0)$$

da cui segue

$$\hat{\Sigma}\vec{w}_2 = \lambda_2\vec{w}_2$$

La dimostrazione prosegue fino a mostrare che $\hat{W} = U_L$

Osservazioni

1) Matrice di correlazione vs matrice di covarianza

Abbiamo lavorato con la decomposizione della matrice di covarianza. Tuttavia è talora preferibile lavorare con la matrice di correlazione che è uniforme rispetto alla scala.

2) Scegliere il numero di "LATENT DIMENSIONS".

Se scegliamo $L = \text{rank}(X)$ otteniamo $\mathcal{L} = 0$. Questa scelta non è ovviamente conveniente. Le strategie tipiche per scegliere L sono:

<u>2a</u> scree plots: un plot degli autovalori λ_j in ordine decrescente. Più aumenta il numero di dimensioni, più gli autovalori diventano piccoli, più il reconstruction error diminuisce.

Definiamo la "fraction of variance explained"

$$F_L = \frac{\sum_{j=1}^{L} \lambda_j}{\sum_{j'=1}^{L_{max}} \lambda_{j'}}$$

e minimizziamo di richiedere $F_L <$ cutoff

<u>2b</u> likelihood profile: per $L < L^*$ (dove L^* denota la dimensione latente vera) il rateo di decrescita di $\mathcal{L}$ è grande, mentre per $L > L^*$ è piccolo. Una maniera per determinare in modo automatico questa variazione è di calcolare la profile likelihood. Supponiamo che λ_L sia la misura dell'errore commesso in un modello di dimensione L tale che

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{L_{max}} \quad \text{(gli autovalori)}$$

Ora separiamo gli autovalori in due gruppi, a seconda che $k < L$ o $k > L$ con L la soglia da determinare. Assumiamo:

$$\begin{cases} \lambda_k \sim N(\mu_1, \sigma^2) & \text{per } k \leq L \\ \lambda_k \sim N(\mu_2, \sigma^2) & \text{per } k > L \end{cases}$$

Facciamo un fit ai λ_k per $k = 1, \dots, L_{max}$ dividendo i dati e calcolando il MLE

$$\mu_1(L) = \frac{\sum_{k \leq L} \lambda_k}{L} \qquad \mu_2(L) = \frac{\sum_{k > L} \lambda_k}{L^{max} - L}$$

$$\sigma^2(L) = \frac{\sum_{k \leq L} (\lambda_k - \mu_1(L))^2 + \sum_{k > L} (\lambda_k - \mu_2(L))^2}{L_{max}}$$

da cui possiamo determinare la log likelihood

$$\ell(L) = \sum_{k=1}^{L} \log N(\lambda_k \mid \mu_1(L), \sigma^2(L)) + \sum_{k=L+1}^{L_{max}} \log N(\lambda_k \mid \mu_2(L), \sigma^2(L))$$

Il picco $L^* = \underset{L}{\operatorname{argmax}}\, \ell(L)$ è ben determinato.