(F94) Metodi statistici per l'apprendimento

Riduzione di dimensionalità

Docente: *Nicolò Cesa-Bianchi* versione 22 maggio 2017

In un problema di apprendimento automatico la scelta degli attributi con i quali rappresentare i dati non è sempre univoca. Per esempio, in un problema di categorizzazione di testi dobbiamo decidere quali parole vogliamo utilizzare fra tutte quelle che compaiono nel dataset. In generale, non tutti gli attributi sono utili ad abbassare il rischio di classificazione. E se forniamo all'algoritmo un gran numero di attributi inutili ci possiamo aspettare un peggioramento delle prestazioni predittive a causa dell'overfitting (il modello da apprendere ha un numero di parametri troppo elevato rispetto alle dimensioni del training set).

Il processo di selezione degli attributi prende il nome di *feature selection* e può essere realizzato utilizzando una varietà di tecniche. Consideriamo istanze $\boldsymbol{x}$ che sono espresse come elementi dello spazio Euclideo a d dimensioni $\mathbb{R}^d$. Dal punto di vista geometrico, la feature selection riduce la dimensionalità dei dati eliminando alcune coordinate, il che corrisponde a proiettare i dati in un sottospazio di $\mathbb{R}^d$ definito da un sottoinsieme dei versori $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d \in \mathbb{R}^d$ della sua base canonica. Più in generale, possiamo ridurre la dimensionalità dei dati proiettandoli in un sottospazio di $\mathbb{R}^d$ definito da una qualsiasi base ortonormale $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathbb{R}^d$ con $k < d$. Questo secondo approccio prende il nome di *feature extraction*. In pratica, come vedremo, la differenza è che in feature selection le nuove coordinate sono un sottoinsieme delle vecchie, mentre in feature extraction le nuove coordinate sono combinazioni lineari delle vecchie.

Una delle tecniche più efficaci di feature extraction è l'analisi delle componenti principali (*Principal Component Analysis*, o PCA). L'idea è la seguente: dato un insieme di istanze $S = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\} \subseteq \mathbb{R}^d$ e un intero $k < d$, vogliamo trovare il sottospazio k-dimensionale di $\mathbb{R}^d$ che approssima meglio S. Formalmente, data una base ortonormale $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ (che definisce un sottospazio k-dimensionale $\mathcal{V} \subseteq \mathbb{R}^d$), indichiamo con $U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_k]$ la matrice con i vettori della base come colonne e indichiamo con $P = UU^\top$ l'operatore lineare che proietta un vettore da $\mathbb{R}^d$ a $\mathcal{V}$. Per capire l'azione di P si consideri il caso $k = 1$. Allora $P = \boldsymbol{u}_1\boldsymbol{u}_1^\top$ e la proiezione di un vettore $\boldsymbol{x}$ nel sottospazio unidimensionale definito da $\boldsymbol{u}_1$ è precisamente $P\boldsymbol{x} = \boldsymbol{u}_1(\boldsymbol{u}_1^\top \boldsymbol{x})$. Si noti più in generale che la trasformazione $\boldsymbol{x} \to U^\top \boldsymbol{x}$ mappa l'istanza $\boldsymbol{x} = (x_1, \dots, x_d)$ nelle k componenti di $\boldsymbol{x}$ sui k elementi della base ortonormale, $U^\top \boldsymbol{x} = (\boldsymbol{u}_1^\top \boldsymbol{x}, \dots, \boldsymbol{u}_k^\top \boldsymbol{x})$. Si noti anche che, se P proietta in uno spazio k-dimensionale $\mathcal{V}$ e $\boldsymbol{x} \in \mathcal{V}$, allora $P\boldsymbol{x} = \boldsymbol{x}$. Ovvero, P è la trasformazione identità per $\mathcal{V}$.

L'errore di approssimazione è quindi calcolato come

$$\sum_{t=1}^{m} \|\boldsymbol{x}_t - P\boldsymbol{x}_t\|^2 \ .$$

Questo misura quanto ciascun punto $\boldsymbol{x}_t$ è distante dalla sua proiezione $P\boldsymbol{x}_t$ nel sottospazio k-dimensionale.

Il problema che risolve PCA è

$$\min_{P \in \mathcal{P}_k} \sum_{t=1}^{m} \|\boldsymbol{x}_t - P\boldsymbol{x}_t\|^2$$

dove $\mathcal{P}_k$ è l'insieme delle matrici di proiezione P su sottospazi k-dimensionali di $\mathbb{R}^d$. Quindi, come preannunciato, le k nuove features $\boldsymbol{u}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{u}_k^\top \boldsymbol{x}$ estratte sono combinazioni lineari delle features precedenti (ovvero, la base in cui i vettori $\boldsymbol{u}_i$ sono espressi).

La norma di Frobenius di una matrice $V = \big[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\big]$ con colonne $\boldsymbol{v}_i \in \mathbb{R}^d$ è definita come

$$\|V\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{d} V_{i,j}^2 = \sum_{i=1}^{m} \|\boldsymbol{v}_i\|^2 \ .$$

Sia $X = \big[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\big]$ la matrice le cui colonne sono gli elementi del dataset S. Allora

$$\sum_{t=1}^{m} \|\boldsymbol{x}_t - P\boldsymbol{x}_t\|_F^2 = \|X - PX\|_F^2 \ .$$

Quindi il problema PCA si riduce ulteriormente a

$$\min_{P \in \mathcal{P}_k} \|X - PX\|_F^2 \ .$$

Per risolvere questo problema, dobbiamo introdurre il concetto di decomposizione sui valori singolari (*Singular Value Decomposition* o SVD) di una matrice.

Sia X una matrice reale $d \times m$. Allora esistono una matrice diagonale $(d \times m)$ $\Sigma = \operatorname{diag}\big(\sigma_1, \ldots, \sigma_r\big)$ con diagonale $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ $(r = \min\{d, m\})$ e due matrici U $(d \times d)$ e V $(m \times m)$ tali che $X = U\,\Sigma\,V^\top$. Gli scalari $\sigma_1, \ldots, \sigma_d$ sono i valori singolari di X. Inoltre le matrici U e V sono ortogonali. Ovvero, $UU^\top = I_d$ e $VV^\top = I_m$, dove I_d $(d \times d)$ e I_m $(m \times m)$ sono matrici identità. Dato che le matrici ortogonali sono anche isometrie, vale la proprietà

$$\|X\|_F = \Big\|U\,\Sigma\,V^\top\Big\| = \|\Sigma\|_F = \sum_{i=1}^{r} \sigma_i^2 \ .$$

Infine, il rango di X è pari al numero dei suoi valori singolari strettamente maggiori di zero.

Iniziamo col dimostrare il seguente risultato: la migliore approssimazione di una matrice X fra tutte le matrici di rango al più k è la matrice ottenuta ponendo a zero i valori singolari più piccoli di σ_k.

Teorema 1 *Sia X una matrice $d \times m$ e sia $U\Sigma V^\top$ la sua SVD con $\Sigma = \operatorname{diag}\big(\sigma_1, \ldots, \sigma_r\big)$. Fissato $1 \leq k \leq r$ e detta $X_k = U\Sigma_k V^\top$ la matrice con $\Sigma_k = \operatorname{diag}\big(\sigma_1, \ldots, \sigma_k, 0 \ldots, 0\big)$, vale*

$$X_k = \operatorname*{argmin}_{Z\,:\,\operatorname{rango}(Z) \leq k} \|X - Z\|_F^2 \ .$$

Dimostrazione. Consideriamo una matrice Z qualunque di dimensioni $d \times m$. Usando le proprietà della SVD, possiamo scrivere

$$\|X - Z\|_F^2 = \Big\|U\Sigma V^\top - UU^\top ZVV^\top\Big\|_F^2 = \Big\|U\big(\Sigma - U^\top ZV\big)V^\top\Big\|_F^2 = \Big\|\Sigma - U^\top ZV\Big\|_F^2$$

Quindi, ricordando che Σ è una matrice diagonale,

$$\left\|\Sigma - U^\top ZV\right\|_F^2 = \sum_{i=1}^{r} \left(\sigma_i - \left(U^\top ZV\right)_{i,i}\right)^2 + \sum_{i\neq j} \left(U^\top ZV\right)_{i,j}^2 .$$

Ora, se $U^\top ZV$ non fosse una matrice diagonale, allora $U^\top ZV = D + S$ dove $D = \operatorname{diag}\left(d_1, \ldots, d_r\right)$ con $d_i = \left(U^\top ZV\right)_{i,i}$ è una matrice diagonale e S non ha tutti gli elementi pari a zero. In questo caso, $Z' = UDV^\top$ sarebbe una migliore approssimazione di Z, in quanto

$$\left\|\Sigma - U^\top Z'V\right\|_F^2 = \|\Sigma - D\|_F^2 = \sum_{i=1}^{r} \left(\sigma_i - \left(U^\top ZV\right)_{i,i}\right)^2 < \left\|\Sigma - U^\top ZV\right\|_F^2 .$$

Allora $U^\top ZV$ dev'essere una matrice diagonale D e, dato che D ha rango k per ipotesi, vale $d_i = 0$ per $i = k+1, \ldots, r$. Quindi

$$\left\|\Sigma - U^\top ZV\right\|_F^2 = \sum_{i=1}^{k} \left(\sigma_i - d_i\right)^2 + \sum_{j=k+1}^{r} \sigma_j^2 .$$

Ora è chiaro che la scelta migliore per D è $d_i = \sigma_i$ per $i = 1, \ldots, k$ e zero altrimenti. Da ciò otteniamo $Z = U \operatorname{diag}\left(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0\right) V^\top = X_k$ con errore di approssimazione

$$\|X - X_k\|_F^2 = \sum_{j=k+1}^{r} \sigma_j^2 .$$

Questo conclude la dimostrazione. □

Ritornando al problema PCA, prima di tutto osserviamo che l'operatore lineare P che proietta il suo argomento in un sottospazio a k dimensioni deve ovviamente avere rango al più k (la dimensione dello span delle colonne di P è al più k per definizione e il rango di P è pari alla dimensione dello span delle colonne). Quindi cerchiamo P di rango al più k tale che $PX = X_k$. Scegliendo $P = U_k U_k^\top$, dove $U_k = \left[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \mathbf{0}, \ldots, \mathbf{0}\right]$ è la matrice le cui prime k colonne sono le stesse della matrice U nell'SVD di $X = U\Sigma V^\top$ otteniamo

$$U_k U_k^\top X = U_k \underbrace{U_k^\top U}_{=I_k} \Sigma V^\top = U_k \Sigma V^\top = U \Sigma_k V^\top$$

dove l'ultima uguaglianza vale perché le colonne di U_k dopo le prime k sono tutte $\mathbf{0}$. Riassumendo, tramite PCA rappresentiamo ciascuno punto $\boldsymbol{x} = (x_1, \ldots, x_d)$ tramite k features $\boldsymbol{x'} = U_k \boldsymbol{x}$ che sono la migliore approssimazione k-dimensionale del punto $\boldsymbol{x}$.