Capitolo 2

Data Preprocessing

2.1 Raccolta di informazioni

2.1.1 Concetti, Proprietà, Misure

2.1.2 Statistiche descrittive

2.2 Tecniche di Pulizia dei dati

2.3 Trasformazioni sui Dati

2.3.1 Lavorare su dati numerici

2.3.2 Scelta di un Campione

2.3.3 Riduzione della Dimensionalità

Un possibile approccio al problema della dimensionalità eccessiva consiste nel ridurre la dimensionalità combinando le features. Le combinazioni lineari dei dati sono particolarmente significative perché sono semplici da calcolare e trattabili analiticamente. Di fatto, tali metodi proiettano i dati multidimensionale in uno spazio con minori dimensioni. In questa sezione tratteremo un approccio particolare, *l'analisi delle componenti principali* (PCA): tale approccio ricerca le proiezioni che meglio rappresentano i dati in termini di errore minimo (espresso con i minimi quadrati. Di fatto, il metodo trova le combinazioni dei dati nei vettori ortonormali pi significativi.

Cominciamo con il fissare gli obiettivi. Consideriamo n vettori $\mathbf{x}_1, \dots \mathbf{x}_n$ tali che $\mathbf{x}_i \in \mathbb{R}^d$. Tali vettori rappresentano le righe del nostro dataset: in pratica, $\mathbf{x}_1^T, \dots, \mathbf{x}_n^T$ sono le righe di una istanza $r \in R(A_1, \dots, A_d)$. In questo caso, d rappresenta la dimensionalità dei nostri dati (il numero di attributi, che vorremmo ridurre), e n rappresenta il numero di istanze: in sintesi, assumiamo $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$. Per semplicità assumiamo che i vettori $\mathbf{x}_i$ siano

normalizzati rispetto alla loro media: ovvero

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \mathbf{0}$$

Il nostro obiettivo è quindi quello di trovare uno spazio $\mathbf{e}_1, \ldots, \mathbf{e}_k$ di vettori ortonormali in $\mathbb{R}^d$ tale che, approssimando $\mathbf{x}_i$ con $\mathbf{y}_i$,

$$\mathbf{x}_i \rightarrow \mathbf{y}_i = \sum_{j=1}^{k} a_{ij}\mathbf{e}_j$$

la funzione di scarto tra i due

$$J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \tag{2.1}$$

sia minimizzata. Si noti che in tal caso, il vettore $\mathbf{a}_i^T = [a_{i1}, \ldots, a_{ik}] \in \mathbb{R}^k$ rappresenta la proiezione di $\mathbf{x}_i$ nello spazio $\mathbf{V} = [\mathbf{e}_1, \ldots, \mathbf{e}_k] \in \mathbb{R}^{d\times k}$. Ritornando al problema originale, possiamo risolvere l'equazione:

$$\begin{aligned}
J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) &= \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} a_{kj}\mathbf{e}_j\|^2 \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}\mathbf{e}_j^T\mathbf{x}_i + \sum_{i=1}^{n} \|\sum_{j=1}^{k} a_{ij}\mathbf{e}_j\|^2 \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}\mathbf{e}_j^T\mathbf{x}_i + \sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}^2\|\mathbf{e}_j\|^2 \\
&\quad +2\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{h=1,h\neq j}^{k} a_{ij}a_{ih}\mathbf{e}_j^T\mathbf{e}_h
\end{aligned}$$

Poiché lo spazio $\mathbf{e}_1, \ldots, \mathbf{e}_k$ deve essere ortonormale, possiamo assumere che $\mathbf{e_i}^T\mathbf{e}_j = 0$ quando $i \neq j$, e $\|\mathbf{e}_i\| = 1$. di conseguenza, la formula di cui sopra diventa:

$$J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) = \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}\mathbf{e}_j^T\mathbf{x}_i + \sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}^2 \tag{2.2}$$

Poiché vogliamo minimizzare tale valore, e stiamo assumendo lo spazio ortonormale fissato, possiamo derivare 2.2 per le componenti a_{ij}, ottenendo

$$\frac{\partial}{\partial a_{ij}} J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) = -2\mathbf{e}_j^T\mathbf{x}_i + 2a_{ij}$$

In pratica, un punto di minimo per la funzione J_k (assumendo $\mathbf{e}_1, \ldots, \mathbf{e}_k$ fissati) si ha con $a_{ij} = \mathbf{e}_j^T \mathbf{x}_i$. Più in generale, dato uno spazio ortonormale $\mathbf{V} = [\mathbf{e}_1, \ldots, \mathbf{e}_k]$, il vettore $\mathbf{a}_i = \mathbf{V}^T \mathbf{x}_i$ rappresenta il vettore $\mathbf{x}_i$ nello spazio $\mathbf{V}$.

Se proviamo a calcolare il valore di J_k utlizzando tali valori ottimali, otteniamo:

$$\begin{aligned} J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij}^2 \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\mathbf{e}_j^T \mathbf{x}_i\right)^T \mathbf{e}_j^T \mathbf{x}_i \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbf{e}_j^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_j \end{aligned}$$

L'ultima uguaglianza può essere ottenuta con banali calcoli algebrici. Si noti che il termine $\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ è proporzionale alla matrice di covarianza (infatti, quest'ultima può essere ottenuta semplicemente moltiplicando per $1/(n-1)$). Poniamo quindi $\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{S}$. Otteniamo:

$$J_k(\mathbf{e}_1, \ldots, \mathbf{e}_k) = \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{j=1}^{k} \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j$$

Riassumendo, dato uno spazio $\mathbf{e}_1, \ldots, \mathbf{e}_k$ di vettori ortonormali, il valore ottimale di 2.1 lo si ottiene con la formula di cui sopra.

Possiamo a questo punto fare una serie di osservazioni. Se fissiamo $\mathbf{e}_1, \ldots, \mathbf{e}_k$ vettori ortonormali, riusciamo a costruire una rappresentazione alternativa di $\mathbf{x}_1, \ldots, \mathbf{x}_n$ combinando linearmente le dimensioni dello spazio ortonormale in maniera ottimale. La questione che ci rimane da risolvere è quindi: qual'è la scelta ottimale per $\mathbf{e}_1, \ldots, \mathbf{e}_k$? Si tratta di risolvere il seguente problema di programmazione vincolata:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{j=1}^{k} \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j \\ \text{subject to} \quad & \|\mathbf{e}_i\| = 1 \qquad i = 1, \ldots, k \\ & \mathbf{e}_i^T \mathbf{e}_j = 0 \qquad i, j = 1, \ldots, k, i \neq j \end{aligned}$$

Tale problema può essere affrontato osservando che la natura dei vincoli ci permette di ricorrere al *metodo del moltiplicatori di Lagrange*. Il metodo dice che, un problema di minizzazione vincolato

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_1(\mathbf{x}) = 0 \\ & \dots \\ & g_n(\mathbf{x}) = 0 \end{aligned}$$

può essere risolto minimizzando la funzione (non vincolata)

$$f'(\mathbf{x}) = f(\mathbf{x}) - \sum_i \lambda_i g_i(\mathbf{x})$$

dove λ_i sono variabili aggiuntive. In pratica, il problema vincolato può essere risolto come problema non vincolato ad una maggiore dimensionalità. Nel nostro caso, possiamo provare a risolvere

$$-\sum_{j=1}^{k} \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j + \sum_{j=1}^{k} \lambda_j (\|\mathbf{e}_j\| - 1)$$

(il termine $\sum_{i=1}^{n} \|\mathbf{x}_i\|^2$) è costante e può essere omesso). derivando rispetto a $\mathbf{e}_j$, e ponendo il risultato uguale a $\mathbf{0}$, otteniamo

$$\mathbf{S}\mathbf{e}_j - \lambda_j \mathbf{e}_j = \mathbf{0}$$

da cui deriva che la funzione J_k ha un punto di minimo quando λ_j è un autovalore di S, e $\mathbf{e_j}$ è il corrispondente autovettore. Notiamo ora che, poiché la matrice S è simmetrica e definita su valori reali, gli autovettori di S sono ortogonali. Questo ci permette di stabilire che, quando $\mathbf{e}_j$ corrisponde ad un autovettore, anche il secondo vincolo è soddisfatto. Se sviluppiamo la relazione $J_k(\mathbf{e}_1, \dots, \mathbf{e}_k)$ assumendo che $\mathbf{e}_1, \dots, \mathbf{e}_k$ siano autovettori distinti di S, otteniamo

$$\begin{aligned} J_k(\mathbf{e}_1, \dots, \mathbf{e}_k) &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{j=1}^{k} \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{j=1}^{k} \lambda_j \mathbf{e}_j^T \mathbf{e}_j \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{j=1}^{k} \lambda_j \end{aligned}$$

Da ciò si deduce che la funzione di cui sopra ha un massimo globale quando $\sum_{j=1}^{k} \lambda_j$ è massimizzata.

Riassumendo, possiamo decomporre una matrice $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ di dati in una nuova matrice $\mathbf{V} = [\mathbf{e}_1, \dots, \mathbf{e}_k]$ dove $\mathbf{e}_i$ è l'autovettore corrispondente all'i-esimo autovalore λ_i (in ordine decrescente), e ogni colonna $\mathbf{x}_i$ è approssimabile

dalla combinazione lineare $\mathbf{y}_i = \sum_{j=1}^{k} a_{ij}\mathbf{e}_j$, dove il vettore $\mathbf{a}_j^T = [a_{1j} \cdots a_{nj}]$ rappresenta la *componente principale* di $\mathbf{X}$ lungo la direzione $\mathbf{e}_j$. Il valore $\sum_{j=1}^{k} \lambda_j$ esprime l'errore commesso nell'approssimare $\mathbf{x}_i$ con $\mathbf{y}_i$: in particolare, se in totale S esibisce p autovalori, il rapporto

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

esprime la misura dell'approssimazione di $\mathbf{X}$ utilizzando k autovettori relativi a k componenti principali.

È interessante notare come la formula $\sum_{j=1}^{k} \lambda_j$ esprima la varianza della matrice $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^T$, mentre

Un esempio: `Iris` data

Consideriamo il seguente campione del dataset `Iris`, così formato:

sepallength	*sepalwidth*	*petallength*	*petalwidth*	*class*
7.2	3.6	6.1	2.5	Iris-virginica
5.9	3	4.2	1.5	Iris-versicolor
5.4	3.4	1.5	0.4	Iris-setosa
5	3.3	1.4	0.2	Iris-setosa
6.7	3	5.2	2.3	Iris-virginica
5.1	3.5	1.4	0.2	Iris-setosa
6.7	3.3	5.7	2.5	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica
5.2	2.7	3.9	1.4	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.1	3.8	1.9	0.4	Iris-setosa
6.6	3	4.4	1.4	Iris-versicolor
5.1	3.5	1.4	0.3	Iris-setosa

Vogliamo effettuare su tale campione un'analisi delle componenti principali. Per prima cosa normalizziamo i dati, utilizzando media e varianza:

$$\begin{aligned} \mu &= [6.0067\ 3.2067\ 3.8867\ 1.4200] \\ \sigma &= [0.8040\ 0.3173\ 1.8291\ 0.9096] \end{aligned}$$

Il risultato è la seguente matrice, che sarà l'input della nostra analisi:

$$\mathbf{X} = \begin{bmatrix} 1.4843 & 1.2397 & 1.2101 & 1.1873 \\ -0.1327 & -0.6514 & 0.1713 & 0.0879 \\ -0.7546 & 0.6093 & -1.3048 & -1.1213 \\ -1.2521 & 0.2942 & -1.3595 & -1.3412 \\ 0.8624 & -0.6514 & 0.7180 & 0.9674 \\ -1.1277 & 0.9245 & -1.3595 & -1.3412 \\ 0.8624 & 0.2942 & 0.9914 & 1.1873 \\ 0.8624 & -0.3362 & 0.9367 & 1.0774 \\ -0.2571 & -1.5969 & 0.6634 & 0.5277 \\ 0.8624 & -0.3362 & 0.9367 & 1.0774 \\ -1.0033 & -1.5969 & 0.0073 & -0.0220 \\ 1.1111 & -0.3362 & 0.5540 & 0.0879 \\ -1.1277 & 1.8701 & -1.0862 & -1.1213 \\ 0.7380 & -0.6514 & 0.2807 & -0.0220 \\ -1.1277 & 0.9245 & -1.3595 & -1.2313 \end{bmatrix}$$

Da $\mathbf{X}$ otteniamo la matrice di correlazione

$$\mathbf{S} = \begin{bmatrix} 1 & -0.2 & 0.89 & 0.87 \\ -0.2 & 1 & -0.47 & -0.43 \\ 0.89 & -0.47 & 1 & 0.98 \\ 0.87 & -0.43 & 0.98 & 1 \end{bmatrix}$$

che esibisce gli autovalori $\lambda_1 = 3.0291, \lambda_2 = 0.8472, \lambda_3 = 0.1108, \lambda_4 = 0.013$. Notiamo che $\sum_{i=1}^{4} = 4$ e che $\lambda_1 + \lambda_2 = 3.8763$ (ovvero, quasi il 97% del totale dei dati). Possiamo quindi approssimare la matrice originaria utilizzando solo gli autovettori relativi ai primi due autovalori:

$$\mathbf{e}_1 = \begin{bmatrix} 0.5197 \\ -0.3017 \\ 0.5696 \\ 0.5607 \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} 0.3754 \\ 0.9207 \\ 0.0541 \\ 0.0926 \end{bmatrix}$$

Utilizzando il nuovo spazio, è possibile definire la trasformazione nelle due principali componenti $\mathbf{X}[\mathbf{e}_1, \mathbf{e}_2]$:

$0.52sl$-$0.302sw$+$0.57pl$+$0.561pw$	$0.375sl$+$0.921sw$+$0.054pl$+$0.093pw$	*class*
1.752354	1.873861	Iris-virginica
0.274451	-0.632121	Iris-versicolor
-1.947978	0.10332	Iris-setosa
-2.265912	-0.396837	Iris-setosa
1.596159	-0.14761	Iris-virginica
-2.391448	0.230147	Iris-setosa
1.589859	0.758106	Iris-virginica
1.687302	0.164595	Iris-virginica
1.021998	-1.481964	Iris-virginica
1.687302	0.164595	Iris-virginica
-0.047758	-1.848442	Iris-versicolor
1.043771	0.145663	Iris-versicolor
-2.397804	1.135858	Iris-setosa
0.727673	-0.309538	Iris-versicolor
-2.32983	0.240319	Iris-setosa

È interessante notare che i dati originali possono essere ricostruiti utilizzando le componenti principali:

$$\mathbf{Y} = \mathbf{A}[\mathbf{e}_1, \mathbf{e}_2]^T = \begin{bmatrix} 1.6141 & 1.1964 & 1.0996 & 1.1559 \\ -0.0946 & -0.6648 & 0.1221 & 0.0954 \\ -0.9736 & 0.6829 & -1.1040 & -1.0826 \\ -1.3266 & 0.3184 & -1.3122 & -1.3072 \\ 0.7742 & -0.6175 & 0.9012 & 0.8813 \\ -1.1565 & 0.9335 & -1.3498 & -1.3195 \\ 1.1109 & 0.2182 & 0.9467 & 0.9616 \\ 0.9387 & -0.3576 & 0.9700 & 0.9613 \\ -0.0251 & -1.6728 & 0.5020 & 0.4359 \\ 0.9387 & -0.3576 & 0.9700 & 0.9613 \\ -0.7187 & -1.6874 & -0.1272 & -0.1979 \\ 0.5971 & -0.1808 & 0.6024 & 0.5987 \\ -0.8198 & 1.7693 & -1.3044 & -1.2393 \\ 0.2620 & -0.5046 & 0.3978 & 0.3793 \\ -1.1207 & 0.9243 & -1.3141 & -1.2840 \end{bmatrix}$$

La trasformazione $\mathbf{Y} = \mathbf{A}[\mathbf{e}_1, \mathbf{e}_2]^T$ restituisce il dataset nello spazio originale, ma con meno rumore rispetto al precedente.

2.3.4 Discretizzazione

Discretizzazione non supervisionata

Discretizzazione supervisionata

2.4 Un caso di studio

Bibliografia

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.