Complementi di Algoritmi e Strutture Dati

Proiezioni casuali

Docente: *Nicolò Cesa-Bianchi* versione 7 aprile 2023

La tecnica del conteggio approssimato permette di stimare in modo efficiente la numerosità degli elementi più frequenti in una collezione. Vediamo ora come la tecnica delle proiezioni casuali ci permette di stimare in modo efficiente le distanze fra coppie di punti nello spazio Euclideo d-dimensionale quando d è grande. Ricordiamo che la distanza Euclidea fra due punti $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ è calcolata come

$$\|\boldsymbol{x} - \boldsymbol{x}'\| = \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2} \ .$$

In molte applicazioni i dati possono essere rappresentati come vettori di numeri. Due esempi importanti sono le immagini (ogni coordinata è un pixel) e i testi (ogni coordinata è una parola del dizionario e il valore della coordinata è la frequenza con la quale la parola compare nel testo). Se consideriamo l'elenco dei film disponibili su Netflix come un dizionario, allora anche un utente di Netflix può essere visto come un vettore di numeri dove ogni coordinata è un film e il valore della coordinata rappresenta una valutazione del film da parte dell'utente.

In tutti questi casi, possiamo interpretare la vicinanza di due punti in $\mathbb{R}^d$ come una misura della similarità fra gli elementi (immagini, testi, utenti) che i punti rappresentano. Quindi, la capacità di calcolare in modo efficiente qual è il punto in un insieme più vicino ad un dato punto (*nearest neighbor*) diventa fondamentale per, ad esempio, suggerire film a nuovi utenti basandosi sui film apprezzati da utenti che hanno un profilo simile (ovvero, le loro codifiche in $\mathbb{R}^d$ sono vicine in termini di distanza Euclidea).

Problema **Nearest neighbor.**
Istanza: Un insieme finito $S \subset \mathbb{R}^d$ e un punto $\boldsymbol{x} \in \mathbb{R}^d$.
Soluzione: $\operatorname*{argmin}_{\boldsymbol{x}' \in S} \|\boldsymbol{x} - \boldsymbol{x}'\|$.

Putroppo, trovare il nearest neighbor in d dimensioni diventa computazionalmente costoso quando $d \gg 1$, come di solito succede nelle applicazioni interessanti. Per esempio, se $|S| = n$ e voglio risolvere il problema nearest neighbor calcolando le distanze fra $\boldsymbol{x}$ e i punti di S impiegherò un tempo dell'ordine di nd. Se devo risolvere il problema ogni volta che aggiungo un nuovo utente ad S impiegherò quindi un tempo dell'ordine di $\sum_{t=1}^{n}(td) = \Theta(n^2 d)$.

Per ovviare a questo problema mostriamo che per ogni $0 < \varepsilon, \delta < 1$ esiste $k = \mathcal{O}\left(\frac{1}{\varepsilon^2} \ln \frac{|S|}{\delta}\right)$ ed esiste una classe $\mathcal{F}$ di funzioni $f : \mathbb{R}^d \to \mathbb{R}^k$ tale che

$$(1-\varepsilon)\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \le \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|^2 \le (1+\varepsilon)\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \qquad \boldsymbol{x}, \boldsymbol{x}' \in S$$

con probabilità almeno $1 - \delta$ rispetto all'estrazione di $f \in \mathcal{F}$.

Per dimostrare questo risultato utilizziamo una tecnica simile al conteggio approssimato. Ovvero, usiamo k funzioni casuali che sono analoghe alle funzioni hash del conteggio approssimato.

Queste funzioni sono rappresentate da k vettori casuali $\boldsymbol{Z}_i, \ldots, \boldsymbol{Z}_k \in \mathbb{R}^d$ estratti in un modo che spiegheremo tra breve. La funzione casuale associata al vettore $\boldsymbol{Z}_j$ è definita come

$$f_j(\boldsymbol{x}) = \boldsymbol{Z}_j^\top \boldsymbol{x} = \sum_{i=1}^{d} Z_{j,i} x_i \ .$$

Il prodotto scalare $\boldsymbol{Z}_j^\top \boldsymbol{x}$ calcola la lunghezza della proiezione di $\boldsymbol{x}$ su $\boldsymbol{Z}_j$ moltiplicata per la lunghezza di $\boldsymbol{Z}_j$. Quindi, usando la funzione f_j possiamo approssimare la distanza $\|\boldsymbol{x} - \boldsymbol{x}'\|$ fra i vettori $\boldsymbol{x}$ e $\boldsymbol{x}'$ con la differenza $\big|f_j(\boldsymbol{x}) - f_j(\boldsymbol{x}')\big|$ fra numeri reali $f_j(\boldsymbol{x})$ e $f_j(\boldsymbol{x}')$. Per ridurre l'errore di approssimazione utilizziamo k funzioni indipendenti invece di una sola.

I vettori $\boldsymbol{Z}_j$ sono ottenuti generando ciascuna componente $Z_{j,i}$ per $i = 1, \ldots, d$ con estrazioni indipendenti da una distribuzione di probabilità con media zero e varianza uno, cioè

$$\mathbb{E}\big[Z_{j,i}\big] = 0 \quad \text{e} \quad \mathrm{Var}\big[Z_{j,i}\big] = 1 \qquad j = 1, \ldots, k \quad i = 1, \ldots, d \ .$$

Quindi,

$$\begin{aligned}
\mathbb{E}\Big[\big(f_j(\boldsymbol{x}) - f_j(\boldsymbol{x}')\big)^2\Big] &= \mathbb{E}\left[\left(\sum_{i=1}^{d}(x_i - x_i')Z_{j,i}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{r=1}^{d}\sum_{s=1}^{d}(x_r - x_r')(x_s - x_s')Z_{j,r}Z_{j,s}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d}(x_i - x_i')^2 Z_{j,i}^2\right] + \mathbb{E}\left[\sum_{r,s\,:\,r\neq s}(x_r - x_r')(x_s - x_s')Z_{j,r}Z_{j,s}\right] \\
&= \sum_{i=1}^{d}(x_i - x_i')^2\,\mathbb{E}\big[Z_{j,i}^2\big] + \sum_{r,s=1}^{d}(x_r - x_r')(x_s - x_s')\mathbb{E}\big[Z_{j,r}\big]\mathbb{E}\big[Z_{j,s}\big] \\
&= \sum_{i=1}^{d}(x_i - x_i')^2\,\mathrm{Var}\big[Z_{j,i}\big] \qquad (1) \\
&= \sum_{i=1}^{d}(x_i - x_i')^2 = \big\|\boldsymbol{x} - \boldsymbol{x}'\big\|^2 \ .
\end{aligned}$$

dove (1) vale perché le $Z_{j,i}$ hanno media zero e quindi

$$\mathrm{Var}[Z_{j,i}] = \mathbb{E}\Big[\big(Z_{j,i} - \mathbb{E}[Z_{j,i}]\big)^2\Big] = \mathbb{E}\big[Z_{j,i}^2\big] \ .$$

Questo dimostra che posso usare $\big(f_j(\boldsymbol{x}) - f_j(\boldsymbol{x}')\big)^2$ per stimare la distanza quadrata $\|\boldsymbol{x} - \boldsymbol{x}'\|^2$.

Definiamo ora la proiezione casuale $f : \mathbb{R}^d \to \mathbb{R}^k$,

$$f(\boldsymbol{x}) = \left(\frac{f_1(\boldsymbol{x})}{\sqrt{k}}, \ldots, \frac{f_k(\boldsymbol{x})}{\sqrt{k}}\right) \ .$$

Si noti che $f(\boldsymbol{x}) = M\boldsymbol{x}$ dove M è la matrice casuale $k \times d$ avente $\boldsymbol{Z}_1/\sqrt{k}, \ldots, \boldsymbol{Z}_k/\sqrt{k}$ come righe. Questo implica che f è una trasformazione lineare, ovvero $f(a\boldsymbol{x} + b\boldsymbol{x}') = af(\boldsymbol{x}) + bf(\boldsymbol{x}')$ per ogni $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ e $a, b \in \mathbb{R}$. Quindi,

$$\mathbb{E}\left[\left\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\right\|^2\right] = \mathbb{E}\left[\left\|f(\boldsymbol{x} - \boldsymbol{x}')\right\|^2\right] = \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}\big[f_j(\boldsymbol{x} - \boldsymbol{x}')^2\big] = \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2$$

dato che ciascun f_j è uno stimatore della distanza quadrata.

Ora, detto $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}'$ e usando sempre il fatto che f è lineare,

$$\frac{\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|^2}{\|\boldsymbol{x} - \boldsymbol{x}'\|^2} = \left\|f\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\right)\right\|^2 \quad .$$

Quindi, se vogliamo dimostrare che

$$(1-\varepsilon)\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2 \leq \left\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\right\|^2 \leq (1+\varepsilon)\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2 \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$$

possiamo equivalentemente dimostrare che $1-\varepsilon \leq \|f(\boldsymbol{v})\|^2 \leq 1+\varepsilon$ per ogni $\boldsymbol{v} \in \mathbb{R}^d$ tale che $\|\boldsymbol{v}\| = 1$.

A questo punto ci serve fare assunzioni sulla distribuzione delle variabili casuali $Z_{j,i}$. Assumiamo quindi che le $Z_{j,i}$ abbiano una distribuzione Normale (ovvero, Gaussiana con media zero e varianza uno). Per le proprietà della Normale, vale che per ogni $\varepsilon, \delta > 0$ fissati e per ogni $\boldsymbol{v} \in \mathbb{R}^d$ di norma unitaria,

$$\mathbb{P}\left(\Big|\, \|f(\boldsymbol{v})\|^2 - 1\Big| > \varepsilon\right) \leq \delta \qquad \text{per} \quad k = \mathcal{O}\left(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta}\right) \tag{2}$$

dove la probabilità è calcolata rispetto all'estrazione delle $\{Z_{j,i} : j = 1, \ldots, k,\ i = 1, \ldots, d\}$.

Si noti che

$$\|f(\boldsymbol{v})\|^2 = \frac{1}{k}\sum_{j=1}^{k}\left(\boldsymbol{Z}_j^\top \boldsymbol{v}\right)^2$$

e inoltre

$$\mathbb{E}\Big[\left(\boldsymbol{Z}_j^\top \boldsymbol{v}\right)^2\Big] = \boldsymbol{v}^\top \mathbb{E}\big[\boldsymbol{Z}_j \boldsymbol{Z}_j^\top\big]\boldsymbol{v} = \boldsymbol{v}^\top I \boldsymbol{v} = \|\boldsymbol{v}\|^2 = 1$$

dove abbiamo usato il fatto che la matrice $M = \mathbb{E}\big[\boldsymbol{Z}_j \boldsymbol{Z}_j^\top\big]$ ha componenti $M_{r,s} = \mathbb{E}\big[Z_{j,r} Z_{j,s}\big]$ tali che

$$M_{r,s} = \begin{cases} 0 & \text{se } r \neq s \\ 1 & \text{altrimenti} \end{cases}$$

Quindi le variabili casuali $V_j = \left(\boldsymbol{Z}_j^\top \boldsymbol{v}\right)^2$ per $j = 1, \ldots, k$ sono i.i.d. con media $\mu = 1$ e (2) può essere riscritta come

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{j=1}^{k} V_j - \mu\right| > \varepsilon\right) \leq e^{-\mathcal{O}(k\varepsilon^2)}$$

Questa diseguaglianza è analoga al Lemma di Chernoff-Hoeffding, con l'unica differenza che qui le V_j non hanno valori limitati. La formula (2) ci dice quindi che un risultato analogo al Lemma di Chernoff-Hoeffding vale anche per variabili casuali del tipo $\left(\boldsymbol{Z}_j^\top \boldsymbol{v}\right)^2$ dove $\boldsymbol{Z}_j$ sono Normali multivariate e $\|\boldsymbol{v}\| = 1$.

Per capire i prossimi passaggi ricordiamo che, per qualsiasi insieme di eventi $A_1, \ldots, A_N$ vale che

$$\mathbb{P}\big(\exists i \,:\, A_i\big) = \mathbb{P}\big(A_1 \cup \cdots \cup A_N\big) \leq \sum_{i=1}^{N} \mathbb{P}(A_i) \ .$$

Nel nostro caso, ci interessano gli eventi

$$A_{\boldsymbol{x},\boldsymbol{x}'} = \left| \frac{\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|^2}{\|\boldsymbol{x} - \boldsymbol{x}'\|^2} - 1 \right| > \varepsilon$$

per ognuna delle $N = \binom{n}{2} \leq n^2$ coppie di punti distinti $\boldsymbol{x}, \boldsymbol{x}' \in S$. Allora, dato un qualunque insieme $S \subset \mathbb{R}^d$ di n punti,

$$\mathbb{P}\left(\exists\, \boldsymbol{x}, \boldsymbol{x}' \in S \,:\, A_{\boldsymbol{x},\boldsymbol{x}'}\right) = \mathbb{P}\left(\bigcup_{\boldsymbol{x},\boldsymbol{x}' \in S} A_{\boldsymbol{x},\boldsymbol{x}'}\right) \leq \sum_{\boldsymbol{x},\boldsymbol{x}' \in S} \mathbb{P}\big(A_{\boldsymbol{x},\boldsymbol{x}'}\big) \leq \sum_{\boldsymbol{x},\boldsymbol{x}' \in S} \delta \leq n^2 \delta$$

per $k = \mathcal{O}\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$.

Da questo ne deduciamo che, per $k = \mathcal{O}\left(\frac{1}{\varepsilon^2} \ln \frac{n}{\delta}\right)$ vale

$$(1-\varepsilon)\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2 \leq \left\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\right\|^2 \leq (1+\varepsilon)\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2 \qquad \text{per ogni } \boldsymbol{x}, \boldsymbol{x}' \in S \tag{3}$$

con probabilità almeno $1 - \delta$ rispetto all'estrazione delle $\{Z_{j,i} \,:\, j = 1, \ldots, k,\ i = 1, \ldots, d\}$.

Se ci accontentiamo di un errore nella stima delle distanze del 10% con probabilità del 99% rispetto all'estrazione di tutte le $Z_{j,i}$, allora ε e δ sono costanti e quindi $k = \mathcal{O}(\log n)$. Il costo per mappare i punti di S in $\mathbb{R}^k$ è $ndk = nd \ln n$ e il costo per calcolare le coppie di distanze fra un $\boldsymbol{x}$ e i punti in S è $n \ln n$. Se devo risolvere il problema nearest neighbor approssimato n volte impiegherò quindi un tempo dell'ordine di $nd \ln n + n^2 \ln n \leq n^2 d$ quando $n = \mathcal{O}(2^d)$.

Se avessimo al più $s < k$ valori non nulli in ciascuna colonna di M, allora il costo per mappare un punto di S in $\mathbb{R}^k$ sarebbe ds. È possibile dimostrare che (3) vale per $k = \mathcal{O}\left(\frac{1}{\varepsilon^2} \ln \frac{n}{\delta}\right)$ e $s = \mathcal{O}\left(\frac{1}{\varepsilon} \ln \frac{n}{\delta}\right)$.