Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων 2020 – 2021

Διδάσκοντες: Α. Παγουρτζής, Θ. Σούλιου, Β. Νάκος

2η Σειρά Ασκήσεων

Άσκηση 1. Στη διάλεξη είδαμε το AMS σκιαγράφημα, το οποίο χρησιμοποείται για να προσεγγίσουμε την ℓ_2 νόρμα ενός διανύσματος σε μία ροή δεδομένων. Σε αυτή την άσκηση θα δούμε μία διαφορετική κατασκευή που έχει ασυμπτωτικά την ίδια επίδοση ως προς χώρο.

Έστω μία τυχαία συνάρτηση κατακερματισμού $h : [n] \to [c/\epsilon^2]$, όπου c μία επαρκώς μεγάλη σταθερά, καθώς και μία συνάρτηση κατακερματισμού $\sigma : [n] \to \{-1, 1\}$. Για κάθε $b \in [c/\epsilon^2]$ κράταμε τον μετρητή

$$C_b := \sum_{i \in [n]: h(i) = b} \sigma(i) \cdot x_i.$$

Έτσι το σκιαγράφημά μας αποτελείται από τις συναρτήσεις h, g και τους μετρητές C. Υποθέτουμε ότι υπάρχει τρόπος να αποθηκεύσουμε τις h και g χρησιμοποιώντας $O(1)$ λέξεις.

α) Εξηγήστε ποιος είναι ο χρόνος ανανέωσης του προαναφερθέντος σκιαγραφήματος. Κοινώς, αν έρθει μία ανανέωση (i, Δ) η οποία προκαλεί $x_i \leftarrow x_i + \Delta$, πόσο γρήγορα μπορούμε να ανανεώσουμε το σκιαγράφημα. Είναι μικρότερος ή μεγαλύτερος από τον χώρο του AMS σκιαγραφήματος όπως το είδαμε στη διάλεξη; Υποθέστε ότι οι h, g μπορούν να αποτιμηθούν σε $O(1)$ χρόνο.

β) Με βάση τους μετρητές C, κατασκευάστε έναν αμερόληπτο εκτιμητή του $\|x\|_2^2$, δηλαδή μία τυχαία μεταβλητή X η οποία ικανοποιεί $\mathbb{E}(X) = \|x\|_2^2$.

γ) Αναλύστε τον προαναφερθέντα εκτιμητή και δείξτε ότι μπορείτε με πιθανότητα $2/3$ να $(1 \pm \epsilon)$-προσεγγίσετε το $\|x\|_2^2$.

Άσκηση 2. Δώστε έναν αλγόριθμο ροής που χρησιμοποιεί $O(\epsilon^{-2} \log n)$ χώρο και με πιθανότητα $2/3$ $(1 \pm \epsilon)$-προσεγγίζει την ποσότητα

$$\sum_{i=1}^{n} (x_i - \mu)^2,$$

όπου $\mu := \frac{1}{n} \sum_{i=1}^{n} x_i$ και $x \in \mathbb{R}^n$ το διάνυσμα το οποίο επιδέχεται αλλαγές. Για την ακρίβεια, ο αλγόριθμός σας θα πρέπει να βρίσκει μία τιμή V ώστε

$$V \in [1 - \epsilon, 1 + \epsilon] \cdot \sum_{i=1}^{n} (x_i - \mu)^2.$$

Άσκηση 3.

Στο πρόβλημα της συσταδοποίησης των k-μέσων η είσοδος αποτελείται από σημεία $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ και έναν θετικό ακέραιο k, και ο σκοπός είναι να βρούμε μία διαμέριση $\mathcal{P}$ του $[n]$ σε k ξένα μεταξύ τους σύνολα, καθώς και $y_1, y_2, \ldots, y_k \in \mathbb{R}^n$ έτσι ώστε να ελαχιστοποιείται η συνάρτηση

$$\text{cost}_{\mathcal{P}}(x_1, x_2, \ldots, x_N) = \sum_{j=1}^{k} \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2.$$

Με απλά λόγια, επιλέγουμε k σημεία $y_1, y_2, \ldots, y_k$ ως κέντρα και ενώνουμε κάθε σημείο x_i σε ένα από τα κέντρα. Η εύρεση της βέλτιστης συσταδοποίησης είναι NP-δύσκολο πρόβλημα, ωστόσο υπάρχουν αποδοτικοί προσεγγιστικοί αλγόριθμοι.

α) Δεδομένης μίας διαμέρισης $\mathcal{P}$ του $[n]$, δείξτε ότι το βέλτιστο σύνολο κέντρων $y_1, y_2, \ldots, y_k$ ικανοποιεί

$$y_j := \frac{1}{|P_j|} \sum_{i \in P_j} x_i, \forall j \in [k].$$

Αυτό υποννοεί ότι μπορούμε να επικεντρωθούμε να βελτιστοποιήσουμε ως προς $\mathcal{P}$.

β) Δείξτε ότι για κάθε $0 < \epsilon < 1/2$ υπάρχει μία γραμμική απεικόνιση $\Pi \in \mathbb{R}^{m \times n}$ με $m = O(\epsilon^{-2} \log N)$ η οποία ικανοποιεί για όλες τις διαμερίσεις $\mathcal{P}$ *ταυτόχρονα*

$$(1-\epsilon)\text{cost}_{\mathcal{P}}(x_1, x_2, \ldots, x_N) \leq \text{cost}_{\mathcal{P}}(\Pi x_1, \Pi x_2, \ldots, \Pi x_N) \leq (1+\epsilon)\text{cost}_{\mathcal{P}}(x_1, x_2, \ldots, x_N)$$

Επιπρόσθετα, η εύρεση του Π μπορεί να γίνει με έναν Monte Carlo πιθανοτικό αλγόριθμο ο οποίος είναι σωστός με μεγάλη πιθανότητα. Άρα αρκεί να λύσουμε το πρόβλημα σε ένα χώρο λογαριθμικής διάστασης ως προς το N.

Προθεσμία υποβολής και οδηγίες. Οι απαντήσεις θα πρέπει να υποβληθούν έως τις 3/6/2021, σε ηλεκτρονική μορφή. Για απορίες / διευκρινίσεις: στείλτε μήνυμα στη διεύθυνση `ads@corelab.ntua.gr`.