

Προηγμένα Θέματα Αλγορίθμων

Πιθανοτικοί Αλγόριθμοι III

Βασίλειος Νάκος

Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστήμιο Αθηνών

Εύρεση Κοντινότερων Γειτόνων

Έστω σύνολο σημείων $\mathcal{P} = \{p_1, \dots, p_n\} \subseteq U$ και ένα σημείο $q \in U$ βρες το

$$p^* = \operatorname{argmin}_{p_i \in \mathcal{P}} d(p_i, q).$$

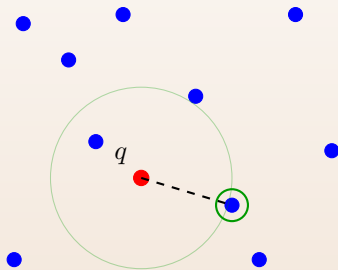
Συνήθως $U = \mathbb{R}^d$ ή $U = \mathbb{R}^d$ και $d(\cdot, \cdot)$ μια συνάρτηση απόστασης.

Εύρεση Κοντινότερων Γειτόνων

Έστω σύνολο σημείων $\mathcal{P} = \{p_1, \dots, p_n\} \subseteq U$ και ένα σημείο $q \in U$ βρες το

$$p^* = \operatorname{argmin}_{p_i \in \mathcal{P}} d(p_i, q).$$

Συνήθως $U = \mathbb{R}^d$ ή $U = \mathbb{R}^d$ και $d(\cdot, \cdot)$ μια συνάρτηση απόστασης.



👉 Επί της ουσίας, πρόβλημα δομών δεδομένων. Θέλουμε να προ-επεξεργαστούμε το $\mathcal{P}$ έτσι ώστε οποιαδήποτε ερώτηση μετά γίνεται πολύ πιο γρήγορη.

Εύρεση Κοντινότερων Γειτόνων

👉 Επί της ουσίας, πρόβλημα δομών δεδομένων. Θέλουμε να προ-επεξεργαστούμε το $\mathcal{P}$ έτσι ώστε οποιαδήποτε ερώτηση μετά γίνεται πολύ πιο γρήγορη.

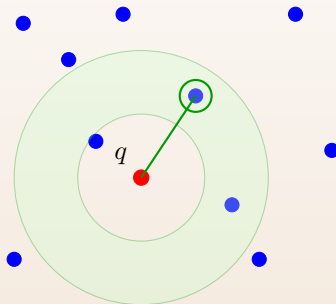
👁 Όλοι οι αλγόριθμοι που γνωρίζουμε έχουν εκθετική εξάρτηση ως προς την διάσταση ... Οπότε, ας προσπαθήσουμε να λύσουμε το εξής πρόβλημα:

c -Προσεγγιστικοί r -Κοντινοί Γείτονες

Έστω σύνολο σημείων $\mathcal{P} = \{p_1, \dots, p_n\} \subseteq U$ και ένα σημείο $q \in U$ ώστε $\min_{p_i \in \mathcal{P}} d(p_i, q) \leq r$. Να βρεθεί οποιοδήποτε σημείο $p_j \in \mathcal{P}$ ώστε $d(p_j, q) \leq cr$.

Εύρεση Κοντινότερων Γειτόνων

Πώς μοιάζουν οπτικά οι c -Προσεγγιστικοί r -Κοντινοί Γείτονες;



Ένα πρόβλημα δομών δεδομένων.

☞ Αυτό που θέλουμε είναι να προεπεξεργαστούμε το σύνολο $\mathcal{P}$ σε μία δομή δεδομένων ώστε μετά να μπορούμε να απαντάμε ερωτήσεις για κάθε q .

Ένα πρόβλημα δομών δεδομένων.

☞ Αυτό που θέλουμε είναι να προεπεξεργαστούμε το σύνολο $\mathcal{P}$ σε μία δομή δεδομένων ώστε μετά να μπορούμε να απαντάμε ερωτήσεις για κάθε q .

Μπορούμε καλύτερα από το *προφανές* χρησιμοποιώντας τυχειότητα;

👁 Σε ένα διαισθητικό επίπεδο, θα μας βόλευε $W : U \rightarrow \mathcal{S}$

👁 Σε ένα διαισθητικό επίπεδο, θα μας βόλευε $W : U \rightarrow \mathcal{S}$ όπου $\mathcal{S}$ ένα μικρότερο σύνολο, και δοθέντων $W(p_1), W(p_2)$ μπορούμε να εξάγουμε αν $d(p_1, p_2) \leq r$ ή $d(p_1, p_2) \geq c \cdot r$.

👁 Σε ένα διαισθητικό επίπεδο, θα μας βόλευε $W : U \rightarrow \mathcal{S}$ όπου $\mathcal{S}$ ένα μικρότερο σύνολο, και δοθέντων $W(p_1), W(p_2)$ μπορούμε να εξάγουμε αν $d(p_1, p_2) \leq r$ ή $d(p_1, p_2) \geq c \cdot r$.

👉 Η συνάρτηση W κάνει ελάττωση διάστασης.

👁 Σε ένα διαισθητικό επίπεδο, θα μας βόλευε $W : U \rightarrow \mathcal{S}$ όπου $\mathcal{S}$ ένα μικρότερο σύνολο, και δοθέντων $W(p_1), W(p_2)$ μπορούμε να εξάγουμε αν $d(p_1, p_2) \leq r$ ή $d(p_1, p_2) \geq c \cdot r$.

👉 Η συνάρτηση W κάνει ελάττωση διάστασης.

$$p = 10000101010101101010110110101010101 \in \{0, 1\}^d$$

$$W(p) = 11101010 \in \{0, 1\}^k$$

Δύο πρώτες λύσεις.

1. Για κάθε $p_i \in \mathcal{P}$ υπολόγισε το $W(p_i)$
2. Για κάθε q υπολόγισε $W(q)$ και σύγκρινε με $W(p_i)$

Δύο πρώτες λύσεις.

1. Για κάθε $p_i \in \mathcal{P}$ υπολόγισε το $W(p_i)$
2. Για κάθε q υπολόγισε $W(q)$ και σύγκρινε με $W(p_i)$

1. Για κάθε $\sigma \in \{0, 1\}^k$ βρες ένα $p_i \in \mathcal{P}$ ώστε $W(p_i)$ **κοντά** σε σ και αποθηκευσέ τα σε έναν πίνακα $A[\sigma]$.
2. Για κάθε q βρες το $W(q) \in \{0, 1\}^k$ και επέστρεψε οποιοδήποτε σημείο στο $A[W(q)]$.

Κατακερματισμός LSH.

Έστω $P_1 > P_2$ αριθμοί στο $[0, 1]$.

Μία LSH συνάρτηση είναι μία **τυχαία** συνάρτηση $h : U \rightarrow \mathcal{S}$ ώστε

- ❶ $d(p, q) \leq r \Rightarrow \Pr[h(p) = h(q)] = P_1$
// αν p, q κοντά τότε θα απεικονιστούν στο ίδιο σημείο με
καλή πιθανότητα

Έστω $P_1 > P_2$ αριθμοί στο $[0, 1]$.

Μία LSH συνάρτηση είναι μία **τυχαία** συνάρτηση $h : U \rightarrow \mathcal{S}$ ώστε

- ❶ $d(p, q) \leq r \Rightarrow \Pr[h(p) = h(q)] = P_1$
//αν p, q κοντά τότε θα απεικονιστούν στο ίδιο σημείο με **καλή** πιθανότητα
- ❷ $d(p, q) > cr \Rightarrow \Pr[h(p) = h(q)] = P_2$
//αν p, q μακριά τότε θα απεικονιστούν στο ίδιο σημείο με **κακή** πιθανότητα

Πώς θα αξιοποιούσαμε τον κατακερματισμό LSH;

👉 LSH: $\rho = \frac{\log(1/P_1)}{\log(1/P_2)}, k := \frac{\log(1/n)}{\log(P_2)}.$

Πώς θα αξιοποιούσαμε τον κατακερματισμό LSH;

👉 LSH: $\rho = \frac{\log(1/P_1)}{\log(1/P_2)}, k := \frac{\log(1/n)}{\log(P_2)}.$

Πάρε k συναρτήσεις LSH και θέσε $g(p) = [h_1(p), h_2(p), \dots, h_k(p)]$.

1. Για κάθε $p_i \in \mathcal{P}$ αποθηκευσέ το στο $A[g(p_i)]$ (ως λίστα) και αποθηκευσέ τα σε έναν πίνακα $A[\sigma]$.
2. Για κάθε q βρες όλα τα $p_i \in A[g(q)]$ και έλεγξε αν $d(p_i, q) \leq cr$.

Πώς θα αξιοποιούσαμε τον κατακερματισμό LSH;

👉 LSH: $\rho = \frac{\log(1/P_1)}{\log(1/P_2)}, k := \frac{\log(1/n)}{\log(P_2)}.$

Πάρε k συναρτήσεις LSH και θέσε $g(p) = [h_1(p), h_2(p), \dots, h_k(p)]$.

1. Για κάθε $p_i \in \mathcal{P}$ αποθηκευσε το στο $A[g(p_i)]$ (ως λίστα) και αποθηκευσε τα σε έναν πίνακα $A[\sigma]$.
2. Για κάθε q βρες όλα τα $p_i \in A[g(q)]$ και έλεγξε αν $d(p_i, q) \leq cr$.

Δεν γνωρίζουμε πως να πετύχουμε ταυτόχρονα υψηλό P_1 και χαμηλό P_2 .

Πώς θα αξιοποιούσαμε τον κατακερματισμό LSH;

👉 LSH: $\rho = \frac{\log(1/P_1)}{\log(1/P_2)}, k := \frac{\log(1/n)}{\log(P_2)}.$

Πάρε k συναρτήσεις LSH και θέσε $g(p) = [h_1(p), h_2(p), \dots, h_k(p)]$.

1. Για κάθε $p_i \in \mathcal{P}$ αποθηκευσέ το στο $A[g(p_i)]$ (ως λίστα) και αποθηκευσέ τα σε έναν πίνακα $A[\sigma]$.
2. Για κάθε q βρες όλα τα $p_i \in A[g(q)]$ και έλεγξε αν $d(p_i, q) \leq cr$.

Δεν γνωρίζουμε πως να πετύχουμε ταυτόχρονα υψηλό P_1 και χαμηλό P_2 .

👉 Αντί για μία, θα χρησιμοποιήσουμε $L := n^\rho$ συναρτήσεις $g_1, g_2, \dots, g_L$.

Πώς θα αξιοποιούσαμε τον κατακερματισμό LSH;

👉 LSH: $\rho = \frac{\log(1/P_1)}{\log(1/P_2)}, k := \frac{\log(1/n)}{\log(P_2)}.$

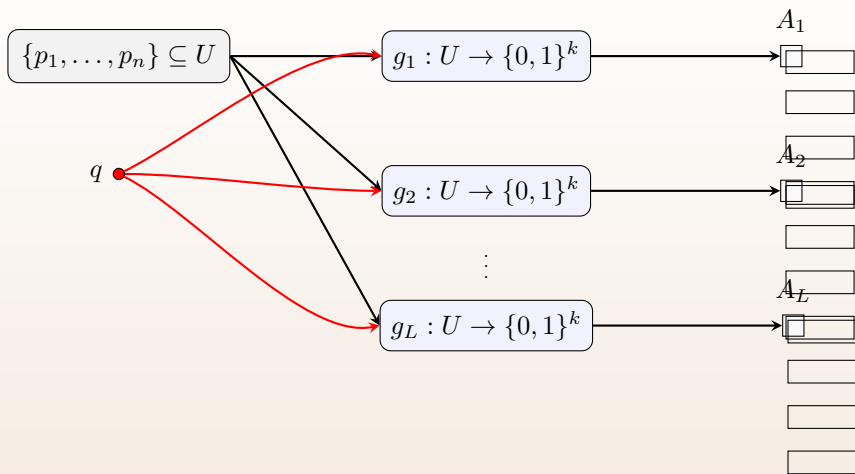
Πάρε k συναρτήσεις LSH και θέσε $g(p) = [h_1(p), h_2(p), \dots, h_k(p)]$.

1. Για κάθε $p_i \in \mathcal{P}$ αποθηκευσέ το στο $A[g(p_i)]$ (ως λίστα) και αποθηκευσέ τα σε έναν πίνακα $A[\sigma]$.
2. Για κάθε q βρες όλα τα $p_i \in A[g(q)]$ και έλεγξε αν $d(p_i, q) \leq cr$.

Δεν γνωρίζουμε πως να πετύχουμε ταυτόχρονα υψηλό P_1 και χαμηλό P_2 .

👉 Αντί για μία, θα χρησιμοποιήσουμε $L := n^\rho$ συναρτήσεις $g_1, g_2, \dots, g_L$.

Εφόσον $\rho < 1$, χρησιμοποιούμε $n^\rho < n$ συναρτήσεις.



Σχήμα: Προσεγγιστικοί Κοντινότεροι Γείτονες με L πίνακες κατακερματισμού και συναρτήσεις $h_\ell : U \rightarrow \{0, 1\}^k$.

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία.

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία. Για κάθε συνάρτηση g_j αποθήκευσε κάθε $p \in \mathcal{P}$ στο $A_j[g_j(p)]$.

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία. Για κάθε συνάρτηση g_j αποθήκευσε κάθε $p \in \mathcal{P}$ στο $A_j[g_j(p)]$. //Χρειάζεσαι να αποθηκεύσεις μόνο n στοιχεία.

👉 Χώρος: $O(L \cdot n + n \cdot d) = O(n^{1+\rho} + nd)$.

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία. Για κάθε συνάρτηση g_j αποθήκευσε κάθε $p \in \mathcal{P}$ στο $A_j[g_j(p)]$. //Χρειάζεσαι να αποθηκεύσεις μόνο n στοιχεία.

👉 Χώρος: $O(L \cdot n + n \cdot d) = O(n^{1+\rho} + nd)$.

Ερώτηση. Δοθέντος q κοίτα στα $A_1[g_1(q)], A_2[g_2(q)], \dots$,

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία. Για κάθε συνάρτηση g_j αποθήκευσε κάθε $p \in \mathcal{P}$ στο $A_j[g_j(p)]$. //Χρειάζεσαι να αποθηκεύσεις μόνο n στοιχεία.

👉 Χώρος: $O(L \cdot n + n \cdot d) = O(n^{1+\rho} + nd)$.

Ερώτηση. Δοθέντος q κοίτα στα $A_1[g_1(q)], A_2[g_2(q)], \dots$, πάρε τα αντίστοιχα p και έλεγξε αν κάποιο ικανοποιεί $\leq cr$. ✓

Χρόνος προεπεξεργασίας και ερώτησης.

👁 Έχουμε παραμέτρους $c, P_1, P_2, \rho := \frac{\log(1/P_1)}{\log(1/P_2)}, L := 2n^\rho$.

L συναρτήσεις $g_1, g_2, \dots, g_L$ και έναν πίνακα A για καθεμία.

Προεπεξεργασία. Για κάθε συνάρτηση g_j αποθήκευσε κάθε $p \in \mathcal{P}$ στο $A_j[g_j(p)]$. //Χρειάζεσαι να αποθηκεύσεις μόνο n στοιχεία.

👉 Χώρος: $O(L \cdot n + n \cdot d) = O(n^{1+\rho} + nd)$.

Ερώτηση. Δοθέντος q κοίτα στα $A_1[g_1(q)], A_2[g_2(q)], \dots$, πάρε τα αντίστοιχα p και έλεγξε αν κάποιο ικανοποιεί $\leq cr$. ✓

👉 Αναμενόμενος χρόνος ανάλογος του πόσα ‘μακρινά’ p_i ‘έκατσαν’ μαζί με το q σε αυτό τον κατακερματισμό.

👁 Ποια η πιθανότητα p και q να ικανοποιούν $g(p) = g(q)$ αν είναι 'κοντα';

👁 Ποια η πιθανότητα p και q να ικανοποιούν $g(p) = g(q)$ αν είναι 'κοντα';

$$\Pr[g(p) = g(q) | d(p, q) \leq r] \leq P_1^k.$$

👁 Ποια η πιθανότητα p και q να ικανοποιούν $g(p) = g(q)$ αν είναι 'κοντά';

$$\Pr[g(p) = g(q) | d(p, q) \leq r] \leq P_1^k.$$

👁 Ποια είναι η πιθανότητα να μην υπάρχει ℓ ώστε $g_\ell(p) = g_\ell(q)$;

👁 Ποια η πιθανότητα p και q να ικανοποιούν $g(p) = g(q)$ αν είναι 'κοντά';

$$\Pr[g(p) = g(q) | d(p, q) \leq r] \leq P_1^k.$$

👁 Ποια είναι η πιθανότητα να μην υπάρχει ℓ ώστε $g_\ell(p) = g_\ell(q)$;

$$\Pr[g_\ell(p) \neq g_\ell(q), \forall \ell \in [L] | d(p, q) \leq r] \leq (1 - P_1^k)^L < 1/3. \checkmark$$

👁 Ποια η πιθανότητα p και q να ικανοποιούν $g(p) = g(q)$ αν είναι 'κοντά';

$$\Pr[g(p) = g(q) | d(p, q) \leq r] \leq P_1^k.$$

👁 Ποια είναι η πιθανότητα να μην υπάρχει ℓ ώστε $g_\ell(p) = g_\ell(q)$;

$$\Pr[g_\ell(p) \neq g_\ell(q), \forall \ell \in [L] | d(p, q) \leq r] \leq (1 - P_1^k)^L < 1/3. \checkmark$$

αν θέσουμε $k := \frac{\log n}{\log(1/P_2)}.$

Ανάλυση Αναμενόμενου Ξρόνου Εκτέλεσης.

Ποιο είναι το αναμενόμενο πλήθος $(\ell, p) \in L \times \mathcal{P}$ ώστε $g_{\ell}(p) = g_{\ell}(q)$ αλλά $d(p, q) < cr$;

Ανάλυση Αναμενόμενου Ξρόνου Εκτέλεσης.

Ποιο είναι το αναμενόμενο πλήθος $(\ell, p) \in L \times \mathcal{P}$ ώστε $g_\ell(p) = g_\ell(q)$ αλλά $d(p, q) < cr$;

☛ Ας ορίσουμε $X_{p,\ell}$ τυχαία μεταβλητή που είναι 1 αν το q με το p ικανοποιούν $g_{\ell,p} = g_{\ell,q}$.

Ανάλυση Αναμενόμενου Ξρόνου Εκτέλεσης.

Ποιο είναι το αναμενόμενο πλήθος $(\ell, p) \in L \times \mathcal{P}$ ώστε $g_{\ell}(p) = g_{\ell}(q)$ αλλά $d(p, q) < cr$;

☛ Ας ορίσουμε $X_{p,\ell}$ τυχαία μεταβλητή που είναι 1 αν το q με το p ικανοποιούν $g_{\ell,p} = g_{\ell,q}$.

👁 Μας ενδιαφέρει το $\sum_{p \in \mathcal{P}} \sum_{\ell \in L} X_{p,\ell}$.

Ανάλυση Αναμενόμενου Ξρόνου Εκτέλεσης.

Ποιο είναι το αναμενόμενο πλήθος $(\ell, p) \in L \times \mathcal{P}$ ώστε $g_\ell(p) = g_\ell(q)$ αλλά $d(p, q) < cr$;

☛ Ας ορίσουμε $X_{p,\ell}$ τυχαία μεταβλητή που είναι 1 αν το q με το p ικανοποιούν $g_{\ell,p} = g_{\ell,q}$.

👁 Μας ενδιαφέρει το $\sum_{p \in \mathcal{P}} \sum_{\ell \in L} X_{p,\ell}$.

$$E[X_{p,\ell}] = \Pr[g_\ell(p) = g_\ell(q) | d(p, q) > r] = P_2^k = \frac{1}{n}.$$

Ανάλυση Αναμενόμενου Ξρόνου Εκτέλεσης.

Ποιο είναι το αναμενόμενο πλήθος $(\ell, p) \in L \times \mathcal{P}$ ώστε $g_\ell(p) = g_\ell(q)$ αλλά $d(p, q) < cr$;

☛ Ας ορίσουμε $X_{p,\ell}$ τυχαία μεταβλητή που είναι 1 αν το q με το p ικανοποιούν $g_{\ell,p} = g_{\ell,q}$.

👁 Μας ενδιαφέρει το $\sum_{p \in \mathcal{P}} \sum_{\ell \in L} X_{p,\ell}$.

$$E[X_{p,\ell}] = \Pr[g_\ell(p) = g_\ell(q) | d(p, q) > r] = P_2^k = \frac{1}{n}.$$

$$\star E \left[\sum_{p \in \mathcal{P}} \sum_{\ell \in L} X_{p,\ell} \right] = \sum_{p \in \mathcal{P}} \sum_{\ell \in L} E[X_{p,\ell}] = n \cdot L \cdot \frac{1}{n} = L = 2n^\rho \checkmark.$$

Οπότε, το πρόβλημα ϵ -προσεγγιστικού Κοντινότερου Γείτονα ανάγεται στο σχεδιασμό LSH.

Μία LSH συνάρτηση είναι μία **τυχαία** συνάρτηση $h : U \rightarrow \{0, 1\}$ ώστε

- ❶ $d(p, q) \leq r \Rightarrow \Pr[h(p) = h(q)] = P_1$
//αν p, q κοντά τότε θα απεικονιστούν στο ίδιο σημείο με **καλή** πιθανότητα

Οπότε, το πρόβλημα c -προσεγγιστικού Κοντινότερου Γείτονα ανάγεται στο σχεδιασμό LSH.

Μία LSH συνάρτηση είναι μία **τυχαία** συνάρτηση $h : U \rightarrow \{0, 1\}$ ώστε

- ❶ $d(p, q) \leq r \Rightarrow \Pr[h(p) = h(q)] = P_1$
//αν p, q κοντά τότε θα απεικονιστούν στο ίδιο σημείο με **καλή** πιθανότητα
- ❷ $d(p, q) > cr \Rightarrow \Pr[h(p) = h(q)] = P_2$
//αν p, q μακριά τότε θα απεικονιστούν στο ίδιο σημείο με **κακή** πιθανότητα

Η περίπτωση $U = \{0, 1\}^d$.

Ας δούμε την περίπτωση $U = \{0, 1\}^d$ και $d(x, y) = \{i : x_i \neq y_i\}$.

1. Έστω $h : \{0, 1\}^d \rightarrow \{0, 1\}$ όπου $h(x)$ ισούται με μία τυχαία συντεταγμένη του x .
2. Διαλέγουμε k τέτοιες τυχαίες συναρτήσεις $h_1, h_2, \dots, h_k$.
3. Θέσε $g(x) = [h_1(x), h_2(x), \dots, h_k(x)]$

Η περίπτωση $U = \{0, 1\}^d$.

Ας δούμε την περίπτωση $U = \{0, 1\}^d$ και $d(x, y) = \{i : x_i \neq y_i\}$.

1. Έστω $h : \{0, 1\}^d \rightarrow \{0, 1\}$ όπου $h(x)$ ισούται με μία τυχαία συντεταγμένη του x .
2. Διαλέγουμε k τέτοιες τυχαίες συναρτήσεις $h_1, h_2, \dots, h_k$.
3. Θέσε $g(x) = [h_1(x), h_2(x), \dots, h_k(x)]$

$$\Pr[h_i(x) = h_i(y)] = 1 - \frac{d(x, y)}{d}, \forall i \in [k]$$

Η περίπτωση $U = \{0, 1\}^d$.

Ας δούμε την περίπτωση $U = \{0, 1\}^d$ και $d(x, y) = \{i : x_i \neq y_i\}$.

1. Έστω $h : \{0, 1\}^d \rightarrow \{0, 1\}$ όπου $h(x)$ ισούται με μία τυχαία συντεταγμένη του x .
2. Διαλέγουμε k τέτοιες τυχαίες συναρτήσεις $h_1, h_2, \dots, h_k$.
3. Θέσε $g(x) = [h_1(x), h_2(x), \dots, h_k(x)]$

$$\Pr[h_i(x) = h_i(y)] = 1 - \frac{d(x, y)}{d}, \forall i \in [k]$$

Άρα

$$\Pr[g(x) = g(y)] = \left(1 - \frac{d(x, y)}{d}\right)^k$$

Η περίπτωση $U = \{0, 1\}^d$.

Ας δούμε την περίπτωση $U = \{0, 1\}^d$ και $d(x, y) = \{i : x_i \neq y_i\}$.

1. Έστω $h : \{0, 1\}^d \rightarrow \{0, 1\}$ όπου $h(x)$ ισούται με μία τυχαία συντεταγμένη του x .
2. Διαλέγουμε k τέτοιες τυχαίες συναρτήσεις $h_1, h_2, \dots, h_k$.
3. Θέσε $g(x) = [h_1(x), h_2(x), \dots, h_k(x)]$

$$\Pr[h_i(x) = h_i(y)] = 1 - \frac{d(x, y)}{d}, \forall i \in [k]$$

Άρα

$$\Pr[g(x) = g(y)] = \left(1 - \frac{d(x, y)}{d}\right)^k$$

Η περίπτωση $U = \{0, 1\}^d$.

① $d(x, y) \leq r \Rightarrow P_1 = 1 - \frac{r}{d} \approx e^{-r/d}$

② $d(x, y) \geq cr \Rightarrow P_2 = 1 - \frac{cr}{d} \approx e^{-c \cdot r/d}$

Η περίπτωση $U = \{0, 1\}^d$.

$$\textcircled{1} \quad d(x, y) \leq r \Rightarrow P_1 = 1 - \frac{r}{d} \approx e^{-r/d}$$

$$\textcircled{2} \quad d(x, y) \geq cr \Rightarrow P_2 = 1 - \frac{cr}{d} \approx e^{-c \cdot r/d}$$

$$\rho = \frac{\log(1/P_1)}{\log(1/P_2)} \approx \frac{1}{c},$$

Η περίπτωση $U = \{0, 1\}^d$.

$$\textcircled{1} \quad d(x, y) \leq r \Rightarrow P_1 = 1 - \frac{r}{d} \approx e^{-r/d}$$

$$\textcircled{2} \quad d(x, y) \geq cr \Rightarrow P_2 = 1 - \frac{cr}{d} \approx e^{-c \cdot r/d}$$

$$\rho = \frac{\log(1/P_1)}{\log(1/P_2)} \approx \frac{1}{c}, \quad k = \frac{\log n}{\log(1/P_2)} = \frac{d \cdot \log n}{r}$$

Για $c = 2$ έχουμε $L \approx n^{\frac{1}{2}}$.

👉 Χώρος: $O(n \cdot d + L \cdot n) = O(n^{\frac{3}{2}} + nd)$.

Η περίπτωση $U = \{0, 1\}^d$.

$$\textcircled{1} \quad d(x, y) \leq r \Rightarrow P_1 = 1 - \frac{r}{d} \approx e^{-r/d}$$

$$\textcircled{2} \quad d(x, y) \geq cr \Rightarrow P_2 = 1 - \frac{cr}{d} \approx e^{-c \cdot r/d}$$

$$\rho = \frac{\log(1/P_1)}{\log(1/P_2)} \approx \frac{1}{c}, \quad k = \frac{\log n}{\log(1/P_2)} = \frac{d \cdot \log n}{r}$$

Για $c = 2$ έχουμε $L \approx n^{\frac{1}{2}}$.

👉 Χώρος: $O(n \cdot d + L \cdot n) = O(n^{\frac{3}{2}} + nd)$.

👉 Χρόνος: $O(L \cdot (k + d)) = O(\sqrt{n} \cdot d \cdot (1 + \frac{\log n}{r}))$.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

👉 Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

☛ Έστω $\Delta = 5$. Δοθέντος $p \in [\Delta]^d$ ας γράψουμε για κάθε συνταγμένη μια συμβολοσειρά μήκους Δ , με τόσους άσσους όσους στην αρχή όσο η τιμή της συντεταγμένης του.

$[1, 4, 2]$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

☛ Έστω $\Delta = 5$. Δοθέντος $p \in [\Delta]^d$ ας γράψουμε για κάθε συνταγμένη μια συμβολοσειρά μήκους Δ , με τόσους άσσους όσους στην αρχή όσο η τιμή της συντεταγμένης του.

$$[1, 4, 2] \Rightarrow [10000, 11110, 11000]$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

☛ Έστω $\Delta = 5$. Δοθέντος $p \in [\Delta]^d$ ας γράψουμε για κάθε συνταγμένη μια συμβολοσειρά μήκους Δ , με τόσους άσσους όσους στην αρχή όσο η τιμή της συντεταγμένης του.

$$[1, 4, 2] \Rightarrow [10000, 11110, 11000] \Rightarrow 100001111011000$$

$$[2, 2, 3]$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

☛ Έστω $\Delta = 5$. Δοθέντος $p \in [\Delta]^d$ ας γράψουμε για κάθε συνταγμένη μια συμβολοσειρά μήκους Δ , με τόσους άσσους όσους στην αρχή όσο η τιμή της συντεταγμένης του.

$$[1, 4, 2] \Rightarrow [10000, 11110, 11000] \Rightarrow 100001111011000$$

$$[2, 2, 3] \Rightarrow [11000, 11000, 11100]$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_1 απόσταση ως $d(x, y) = \sum_{i=1}^d |x_i - y_i|$.

☛ Μπορούμε να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων σχεδιάζοντας ένα σχήμα LSH.

Μπορούμε όμως και να απεικονίσουμε κάθε $p \in [\Delta]^d$ στο $\{0, 1\}^{???}$ με την απόσταση για την οποία έχουμε ήδη γνωστές δομές δεδομένων και χρησιμοποιήσουμε το εκεί αποτέλεσμα.

☛ Έστω $\Delta = 5$. Δοθέντος $p \in [\Delta]^d$ ας γράψουμε για κάθε συνταγμένη μια συμβολοσειρά μήκους Δ , με τόσους άσσους όσους στην αρχή όσο η τιμή της συντεταγμένης του.

$$[1, 4, 2] \Rightarrow [10000, 11110, 11000] \Rightarrow 100001111011000$$

$$[2, 2, 3] \Rightarrow [11000, 11000, 11100] \Rightarrow 100001111011000$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω οι αντίστοιχες απεικόνιση $f : \Delta \Rightarrow \{0, 1\}$ και $e : [\Delta]^d \Rightarrow \{0, 1\}^{\Delta \cdot d}$

👁 Ισχύει ότι $d(f(x), f(y)) = |x - y|$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω οι αντίστοιχες απεικόνιση $f : \Delta \Rightarrow \{0, 1\}$ και $e : [\Delta]^d \Rightarrow \{0, 1\}^{\Delta \cdot d}$

👁 Ισχύει ότι $d(f(x), f(y)) = |x - y|$

👁 Ισχύει ότι $|d(e(x), e(y))| = \|x - y\|_1, \forall x, y \in [\Delta]^d$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω οι αντίστοιχες απεικόνιση $f : \Delta \Rightarrow \{0, 1\}$ και $e : [\Delta]^d \Rightarrow \{0, 1\}^{\Delta \cdot d}$

👁 Ισχύει ότι $d(f(x), f(y)) = |x - y|$

👁 Ισχύει ότι $|d(e(x), e(y))| = \|x - y\|_1, \forall x, y \in [\Delta]^d$

Μπορούμε να χρησιμοποιήσουμε την γνωστή δομή από πριν και να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων:

👉 Χώρος: $O(n \cdot d + L \cdot n) = O(n^{\frac{3}{2}} + n \cdot \Delta d)$.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω οι αντίστοιχες απεικόνιση $f : \Delta \Rightarrow \{0, 1\}$ και $e : [\Delta]^d \Rightarrow \{0, 1\}^{\Delta \cdot d}$

👁 Ισχύει ότι $d(f(x), f(y)) = |x - y|$

👁 Ισχύει ότι $|d(e(x), e(y))| = \|x - y\|_1, \forall x, y \in [\Delta]^d$

Μπορούμε να χρησιμοποιήσουμε την γνωστή δομή από πριν και να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων:

👉 Χώρος: $O(n \cdot d + L \cdot n) = O(n^{\frac{3}{2}} + n \cdot \Delta d)$.

👉 Χρόνος: $O(L \cdot (k + d)) = O(\sqrt{n} \cdot \Delta \cdot d \cdot (1 + \frac{\log n}{r}))$.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_1 απόσταση.

Έστω οι αντίστοιχες απεικόνιση $f : \Delta \Rightarrow \{0, 1\}$ και $e : [\Delta]^d \Rightarrow \{0, 1\}^{\Delta \cdot d}$

👁 Ισχύει ότι $d(f(x), f(y)) = |x - y|$

👁 Ισχύει ότι $|d(e(x), e(y))| = \|x - y\|_1, \forall x, y \in [\Delta]^d$

Μπορούμε να χρησιμοποιήσουμε την γνωστή δομή από πριν και να λύσουμε το πρόβλημα των προσεγγιστικών γειτόνων:

👉 Χώρος: $O(n \cdot d + L \cdot n) = O(n^{\frac{3}{2}} + n \cdot \Delta d)$.

👉 Χρόνος: $O(L \cdot (k + d)) = O(\sqrt{n} \cdot \Delta \cdot d \cdot (1 + \frac{\log n}{r}))$.

Λύνεται και καλύτερα αλλά δεν θα το δούμε.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_2 απόσταση ως

$$d(x, y) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}.$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_2 απόσταση ως

$$d(x, y) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}.$$

Ας υποθέσουμε για απλοποίηση πως κάθε $p_i \in \mathcal{P}$ έχει μέτρο 1, δηλαδή $\|p_i\|_2 = 1$.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.

Έστω $\mathcal{P} \subseteq [\Delta]^d$. Ορίζουμε την ℓ_2 απόσταση ως

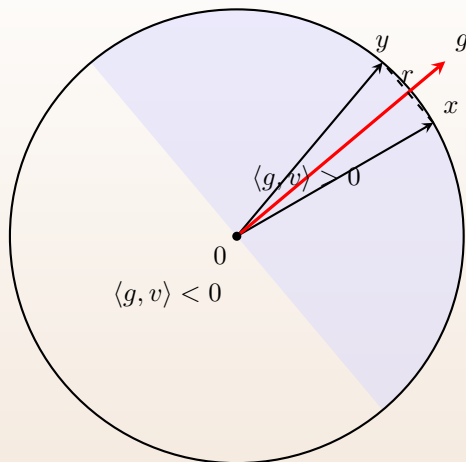
$$d(x, y) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}.$$

Ας υποθέσουμε για απλοποίηση πως κάθε $p_i \in \mathcal{P}$ έχει μέτρο 1, δηλαδή $\|p_i\|_2 = 1$.

Ας πάρουμε ανεξάρτητες τυχαίες μεταβλητές $g_1, g_2, \dots, g_d \sim \mathcal{N}(0, 1)$ και ας ορίσουμε

$$h(x) = \text{sign}(g_1 x_1 + g_2 x_2 + \dots + g_d x_d) = \text{sign}(\langle g, x \rangle).$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.



Σχήμα: Η συνάρτηση $h(x) = \text{sign}(\langle g, x \rangle)$ είναι σαν να επιλέγει ένα τυχαίο (υπερ)-επίπεδο και βάζει στο x ένα bit που δείχνει αν είναι από την μία ή από την άλλη πλευρά του (υπερ)-επιπέδου.

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.

👁 Η πιθανότητα x, y να καταλήξουν στο ίδιο υπερεπίπεδο είναι $1 - \frac{2}{\pi}\alpha$ με $\alpha := \sin^{-1}(r/2)$. Οπότε

$$\rho = \frac{\log(1 - \frac{2}{\pi}\sin^{-1}(r/2))}{\log(1 - \sin^{-1}(cr/2))} \leq \frac{1}{c}.$$

Η περίπτωση $U = \mathbb{R}^d$ με ℓ_2 απόσταση.

👁 Η πιθανότητα x, y να καταλήξουν στο ίδιο υπερεπίπεδο είναι $1 - \frac{2}{\pi}\alpha$ με $\alpha := \sin^{-1}(r/2)$. Οπότε

$$\rho = \frac{\log(1 - \frac{2}{\pi}\sin^{-1}(r/2))}{\log(1 - \sin^{-1}(cr/2))} \leq \frac{1}{c}.$$

Αν πάρουμε $r = 1/10$ (ή μία σταθερά) μπορούμε να πάρουμε μία δομή για c -προσεγγιστικούς γείτονες με τις ίδιες προδιαγραφές με πριν.