CIS 700: “algorithms for Big Data”

Lecture 11: K-means

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

K-means Clustering

- Given $\mathrm{X} = \{x_1, \dots, x_n\} \in \mathbb{R}^d$ find a set of centers $C = (c_1, \dots, c_k)$ that minimizes

$$\sum_{x \in X} \min_{i \in [k]} ||x - c_i||^2$$

- NP-hard problem
- Popular heuristic local search (Lloyd's alg.)
- For a fixed partitioning $P_1, \dots, P_k$:

$$c_j = \frac{1}{|P_j|} \cdot \sum_{i \in P_j} x_i$$

Dimension reduction for K-means

- Let $cost_P(X) = \inf_c cost_{P,c}(X)$
- For $0 < \epsilon < \frac{1}{2}$ let $f: X \to \mathbb{R}^n$ be such that

$$\forall i,j: (1-\epsilon)\left|\left|x_i - x_j\right|\right|_2^2 \leq \left|\left|f(x_i) - f(x_j)\right|\right|_2^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|_2^2$$

- $\hat{P}$ is a γ-approx. clustering for $f(X)$
- P^* is an optimal clustering for X
- **Lemma**.

$$cost_{\hat{P}} \leq \gamma\left(\frac{1+\epsilon}{1-\epsilon}\right) cost_{P^*}(X)$$

Dimension reduction for K-means

- Let $cost_P(X) = \inf_c cost_{P,c}(X)$
- For $0 < \epsilon < \frac{1}{2}$ let $f: X \to \mathbb{R}^{d'}$ be such that
$\forall i,j: (1-\epsilon)\left|\left|x_i - x_j\right|\right|_2^2 \leq \left|\left|f(x_i) - f(x_j)\right|\right|_2^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|_2^2$
- $\hat{P}$ is a γ-approx. clustering for $f(X)$
- P^* is an optimal clustering for X
- **Lemma**.

$$cost_{\hat{P}} \leq \gamma\left(\frac{1+\epsilon}{1-\epsilon}\right) cost_{P^*}(X)$$

- $d' = O\left(\log\frac{n}{\epsilon^2}\right)$ suffices by the JL-lemma

Dimension reduction for K-means

- Fix a partition $P = (P_1, \dots, P_k)$

$$cost_P(X) = \sum_{j\in[k]} \sum_{i\in P_j} \left\| x_i - \frac{1}{|P_j|} \sum_{i'\in P_j} x_{i'} \right\|_2^2$$

$$= \sum_{j\in[k]} \frac{1}{|P_j|} \sum_{i\in P_j} \left(\sum_{i'\in P_j} ||x_i||_2^2 - 2\langle x_i, \sum_{i'\in \mathrm{P}_j} x_i' \rangle + \left\| \sum_{i'\in \mathrm{P}_j} x_i' \right\|_2^2 \right)$$

$$= \sum_{j\in[k]} \frac{1}{|P_j|} \sum_{i\in P_j} \sum_{i'\in P_j} \left(\frac{||x_i||_2^2 + ||x_i'||_2^2}{2} - \langle x_i, x_{i'} \rangle \right)$$

$$\sum_{j\in[k]} \frac{1}{2|P_j|} \sum_{i\in P_j} \sum_{i'\in P_j} \left(||x_i - x_{i'}||_2^2 \right)$$

- $(1-\epsilon)cost_P(X) \le cost_P\big(f(X)\big) \le (1+\epsilon)cost_P(X)$
- $(1-\epsilon)cost_{\hat{P}}(X) \le cost_{\hat{P}}\big(f(X)\big) \le \gamma\ cost_{P^*}\big(f(X)\big) \le \gamma\ cost_{P^*}(X)$

K-means++ Algorithm

- First center uniformly at random from X
- For a set of centers C let:

$$d^2(x, C) = \min_{c \in C} \left|\left|x - c\right|\right|_2^2$$

- Fix current set of centers C
- Subsequent centers: each x_i with prob.

$$\frac{d^2(x_i, C)}{\sum_{x_j \in X} d^2(x_j, C)}$$

- Gives $O(\log k)$-approx. to OPT in expectation

K-means|| Algorithm

- First center C: sample a point uniformly
- Initial cost $\psi = \sum_x d^2(x, C)$
- For $O(\log \psi)$ times do:
  - Repeat ℓ times (in parallel)
    - $C' =$ sample each $x_i \in X$ indep. with prob.

$$p_x = \frac{d^2(x_i, C)}{\sum_{x_j \in X} d^2(x_j, C)}$$

    - $C \leftarrow C \cup C'$
- For $x \in C$:

  w_x = the #points belonging to this center
- Cluster the weighted points in C into k clusters

K-means‖ Algorithm

- Oversampling factor $\ell = \Theta(k)$
- #points in C: $\ell \log \psi$
- **Thm.** If α**-**approx. used in the last step then k-means‖ obtains an $O(\alpha)$-approx. to k-means
- If Ψ and Ψ' are the costs of clustering before and after one outer loop iteration then:

$$E[\Psi'] = O(OPT) + \frac{k}{e\ell}\Psi$$

K-means|| Analysis

- For a set of points $A = \{a_1, \ldots, a_t\}$ *centroid* c_A*:*

$$c_A = \frac{1}{|T|} \sum a_t$$

- Order $a_1, \ldots, a_T$ in the increasing order by distance from c_A
- Fix a cluster A in OPT
- Fix C prior to the iteration and let:

$$\phi(C) = \sum_x d^2(x, C)$$

$$\phi_A(C) = \sum_a d^2(a, C)$$

- Let $p_t = \frac{d^2(a_t, C)}{\phi(C)}$ be the probability of selecting a_t
- Probability that a_t is the smallest one chosen:

$$q_t = p_t \prod_{j=1}^{t-1} (1 - p_j)$$

K-means‖ Analysis

- Can either assign all points to some selected a_t or keep the original clustering:

$$s_t = \min\left(\phi_A, \sum_{a\in A} ||a - a_t||^2\right)$$

- $E[\phi_A(C \cup C')] \leq \sum_t q_t s_t + q_{T+1}\, \phi_A(C)$

where q_{T+1}= prob. that no point in A is selected

- Simplifying assumption: consider the case when all $p_t = p$ (mean field analysis)
- $q_t = p(1-p)^t$ (decreasing sequence)

K-means|| Analysis

- $s'_t = \sum_{a \in A} ||a - a_t||^2$
- $\{s'_t\}$ is an increasing sequence

$$\sum_t q_t s_t \le \sum_t q_t s'_t$$
$$\le \frac{1}{T}\left(\sum_t q_t \sum_t s'_t\right)$$
$$= \left(\sum_t q_t \cdot \frac{1}{T}\sum_t s'_t\right)$$
$$= \left(\sum_t q_t \cdot 2\,\phi_A^*\right)$$

- $E[\phi_A(C \cup C')] \le (1 - q_{T+1})\, 2\,\phi_A^* + q_{T+1}\,\phi_A(C)$