Parallel *k*-Means in Theory in Practice

Florian Schoppmann

August 1, 2012

k-Means

Input:

- Number of desired means $k \in \mathbb{N}$
- Set of points $P \subset \mathbb{R}^d$ (or multiset)

Output:

- Set of k means $C = \{c_1, \ldots, c_k\}$

Lloyd's Heuristic

- Additional input: Seeding C

1: **repeat**
2: **for** $x \in P$ **do**
3: $a[x] \leftarrow \arg\min_{c \in C} \operatorname{dist}(x, c)$
4: **for** $c \in C$ **do**
5: $c \leftarrow \operatorname{mean}(\{x \in P \mid a[x] = c\})$
6: **until** C did not change in last iteration

Lloyd's Heuristic

1: $C \leftarrow \{\text{random } p \in P\}$
2: **while** $|C| < k$ **do**
3: $\quad C \leftarrow C \cup \{\text{random } p \in P, \text{sampled} \sim \text{dist}(p, C)^2\}$

Implementation in SQL (1/2)

$C \leftarrow \{\text{random } p \in P\}$:

```
1 SELECT ARRAY[(
2     SELECT CAST($expr_point AS DOUBLE PRECISION[])
3     FROM $rel_source
4     WHERE $expr_id = (
5         SELECT weighted_sample($expr_id, 1)
6         FROM $rel_source
7     )
8 )]
```

Implementation in SQL (2/2)

$C \leftarrow C \cup \{\text{random } p \in P, \text{sampled} \sim \text{dist}(p, C)^2\}$

```
1  SELECT centroids || $expr_point
2  FROM $rel_source
3  WHERE $expr_id = (
4      SELECT weighted_sample(
5          $expr_id, (
6          closest_column(
7              centroids,
8              $expr_point,
9              fn_squared_dist
10         )).distance
11     ) FROM $rel_source)
```

k-means||

1: $C \leftarrow \{\text{random } p \in P\}$
2: $\Phi_0 \leftarrow \Phi(C)$
3: **for** $O(\log \Phi_0)$ times **do**
4: $\quad C' \leftarrow \{p \in P, \text{ each with prob. } \frac{\ell \cdot \text{dist}(p,C)^2}{\Phi(C)}\}$
5: $\quad C \leftarrow C \cup C'$
6: **for** $c \in C$ **do**
7: $\quad w_c \leftarrow \#$ points in P "assigned" to c
8: Run (weighted) k-means++ on C