Clustering: I

Piotr Indyk

Clustering

- Problem:
  - Given: points $p^1 \dots p^n$
  - Goal: partition the points into clusters $C^1 \dots C^k$ in a "meaningful" way
- Fundamental tool in data analysis
- Hard for
  - Massive data (i.e., large n)
  - High dimensional data

Menu

- k-means: fast heuristic, often used in practice
- k-center: provable, useful algorithm
- agglomerative clustering

k-means

- Goal: find *means* $c^1 \ldots c^k$ and partitioning $C^1 \ldots C^k$ which minimize

$$\sum_{j=1}^{k} \sum_{p \in C^j} \|p - c^j\|_2^2$$

- Idea:
  - If we knew the means c^j , we would assign each p to closest c^j
  - If we knew the clusters C^j, we would define c^j to be the best mean for C^j

- Algorithm:
  - Make an initial choice of $c^1 \ldots c^k$ (e.g., sampling)
  - Alternate the above steps

Finding Best c for C

Not surprisingly, $c = \frac{\sum_{p \in C} p}{|C|}$.

Proof:

- Want to minimize

$$
\begin{aligned}
\sum_{p \in C} \|p - c\|_2^2 &= \sum_{p \in C} \sum_{j=1}^{d} |p_j - c_j|^2 \\
&= \sum_{j=1}^{d} \sum_{p \in C} |p_j - c_j|^2 \\
&= \sum_{j=1}^{d} \sum_{p \in C} [p_j^2 - 2p_j c_j + c_j^2] \\
&= \sum_{j=1}^{d} [\sum_{p \in C} p_j^2 - 2 \sum_{p \in C} p_j c_j + |C| c_j^2]
\end{aligned}
$$

- Can optimize each c_j separately

- The quadratic function is minimized if the derivative is equal to 0

k-means

- Often works well in practice
- No theoretical guarantees (at least not for this version of the algorithm)
- Running time: $O(kn)$ per iteration

Expectation Maximization

- Problem with k-means: each point can belong to only one cluster
- Relaxation: for each p, C, we have $\Pr[p \in C]$
- Compute the mean by taking

$$c = \frac{\sum_p p \Pr[p \in C]}{\sum_p \Pr[p \in C]}$$

- Few theoretical guarantees
- Used everywhere
- Original paper: most cited CS paper (according to NEC CiteSeer)
- Running time: $O(kn)$ per iteration

k-center

- Goal: find *centers* $c^1 \ldots c^k$ and partitioning $C^1 \ldots C^k$ which minimize

$$\max_j \max_{p \in C^j} \|p - c^j\|$$

- As before, each C^j consists of points for which c^j is the closest center

- Can be formulated in any *metric space*, i.e., for any distance function $D(\cdot, \cdot)$ that is:
  - Reflexive: $D(p, q) = 0$ iff $p = q$
  - Symmetric: $D(p, q) = D(q, p)$
  - Satisfies triangle inequality:
    $D(p, q) + D(q, s) \geq D(p, s)$

- Will show a 2-approximation algorithm for this problem

Greedy Algorithm for k-center

- "Guess" the minimum value R of the objective function

- Repeat:
  - Pick any remaining point as the next center $\tilde{c}$
  - Define $\tilde{C}$ to be the set of all points within distance $2R$ from $\tilde{c}$
  - Remove $\tilde{C}$

- Correctness:
  - Let C^j be the opt cluster containing $\tilde{c}$
  - By triangle ineq., $C^j \subset \tilde{C}$, since for any $p \in C^j$

$$D(p, \tilde{c}) \leq D(p, c^j) + D(c^j, \tilde{c}) \leq 2R$$

  - I.e., we remove at least one cluster at a time
  - The total number of iterations is $\leq k$

Analysis ctd.

- Guess: implemented by trying $R = 1, (1+\epsilon), (1+\epsilon)^2 \ldots$
- Alternatively, we can try $R = D(p, q)$ for all p, q
- Modulo that, the running time is $O(kn)$
- In practice, the choice of k is as arbitrary as the choice of R

Hierarchical Agglomerative Clustering

- Start with clusters $C^i = \{p^i\}$, $i = 1 \ldots n$
- Repeat:
  - Pick a pair of clusters C^i, C^j which minimizes $D(C^i, C^j)$
  - Replace C^i, C^j by $C^i \cup C^j$
- What is $D(C, C')$?
  - S-LINK: $\min_{p \in C, q \in C'} D(p, q)$
  - C-LINK: $\max_{p \in C, q \in C'} D(p, q)$
  - A-LINK: $\frac{\sum_{p \in C, q \in C'} D(p,q)}{|C||C'|}$
  - YOP-LINK: your own personal link
- Running time: $\Omega(n^2)$

S-LINK Deja Vu

More clustering

Avrim Blum, "Correlation Clustering"

This Wednesday (October 2), 4:15pm, NE43-518