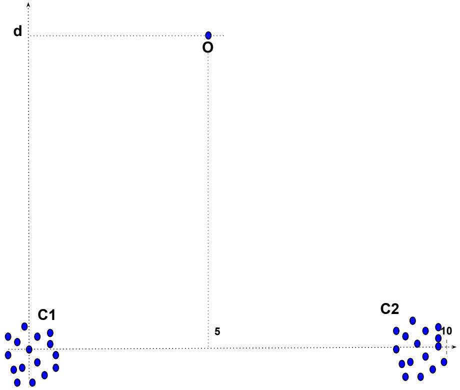
K-Means++: Deterministic vs. Random[2]

Mauro Sozio

April 7, 2017

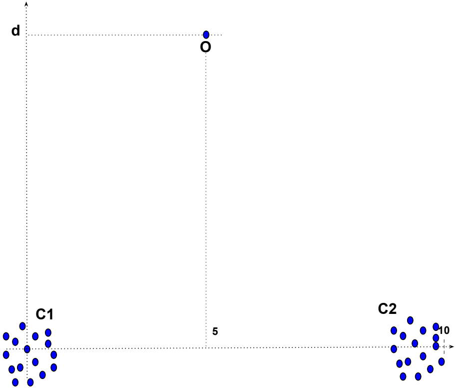
[2] Advanced topic, not asked at the exam

Input Points



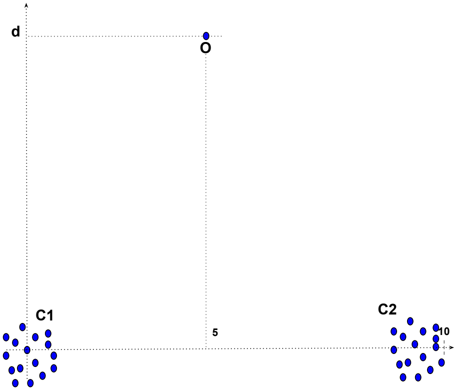
- $C1$ and $C2$ contain $\frac{n}{2}$ points each. $O = (5, d)$ (d large constant).
- distance between any two points in $C1$ and any two points in $C2$ is at most 1.
- distance between any point in $C1$ and O, and any point in $C2$ and O is $\approx d$ (d large).
- We study the SSE of the algorithms as a function of n with d being a large constant.

Deterministic Algorithm ($k = 2$)



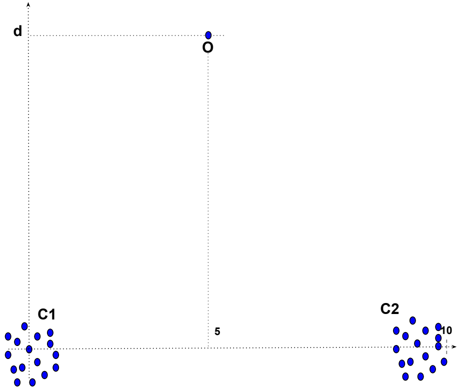
- Selects O and one point in $C1$ or $C2$ as centroids.
- SSE $\approx \frac{n}{2} + \frac{n}{2} \cdot d^2$. Huge error!

k-Means++ ($k = 2$)



- Suppose we select one point p in $C1$ as first centroid. Then, $\Phi_X \approx 10^2 \cdot \frac{n}{2} + d^2$.
- Any given point in $C2$ is selected with prob. $\approx \frac{100}{\Phi_X}$. The prob. that one point in $C2$ is selected as second centroid is $\approx \frac{n}{2} \cdot \frac{100}{\Phi_X}$, while O is selected with prob. $\approx \frac{d^2}{\Phi_X}$.

k-Means++ ($k = 2$)



- O is selected with probability at most $\frac{1}{n} + \frac{d^2}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + d^2 \cdot \frac{n}{2}$.
- two points in a same cluster ($C1$ or $C2$) are selected with probability at most $\frac{1}{\Phi_X}$ in which case the SSE is $\approx \frac{n}{2} + 100 \cdot \frac{n}{2} + d^2$.
- one point in $C1$ and one point in $C2$ are selected as centroids with probability $1 - \frac{1}{n} - \frac{d^2}{\Phi_X} - \frac{1}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + \frac{n}{2} + d^2$.

k-Means++: SSE on average

Therefore, we can compute the SSE on average for k-Means++ as follows. Remember:

- $\Phi_X \approx 10^2 \cdot \frac{n}{2} + d^2$.
- O is selected with probability at most $\frac{1}{n} + \frac{d^2}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + d^2 \cdot \frac{n}{2}$.
- two points in a same cluster ($C1$ or $C2$) are selected with probability at most $\frac{1}{\Phi_X}$ in which case the SSE is $\approx \frac{n}{2} + 100 \cdot \frac{n}{2} + d^2$.
- one point in $C1$ and one point in $C2$ are selected as centroids with probability $1 - \frac{1}{n} - \frac{d^2}{\Phi_X} - \frac{1}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + \frac{n}{2} + d^2$.

k-Means++: SSE on average

Therefore, we can compute the SSE on average for k-Means++ as follows. Remember:

- $\Phi_X \approx 10^2 \cdot \frac{n}{2} + d^2$.
- O is selected with probability at most $\frac{1}{n} + \frac{d^2}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + d^2 \cdot \frac{n}{2}$.
- two points in a same cluster ($C1$ or $C2$) are selected with probability at most $\frac{1}{\Phi_X}$ in which case the SSE is $\approx \frac{n}{2} + 100 \cdot \frac{n}{2} + d^2$.
- one point in $C1$ and one point in $C2$ are selected as centroids with probability $1 - \frac{1}{n} - \frac{d^2}{\Phi_X} - \frac{1}{\Phi_X}$, in which case the SSE is $\approx \frac{n}{2} + \frac{n}{2} + d^2$.

It follows that:

$$
\begin{aligned}
\text{Expected SSE} \approx & \left(\frac{1}{n} + \frac{d^2}{\Phi_X}\right) \cdot \left(\frac{n}{2} + d^2 \cdot \frac{n}{2}\right) + \\
& \frac{1}{\Phi_X} \cdot \left(\frac{n}{2} + 100 \cdot \frac{n}{2} + d^2\right) + \\
& \left(1 - \frac{1}{n} - \frac{d^2}{\Phi_X} - \frac{1}{\Phi_X}\right) \cdot \left(\frac{n}{2} + \frac{n}{2} + d^2\right)
\end{aligned}
$$

[8] Advanced topic, not asked at the exam

Last steps:

Remember: $\Phi_X \approx 10^2 \cdot \frac{n}{2} + d^2$.

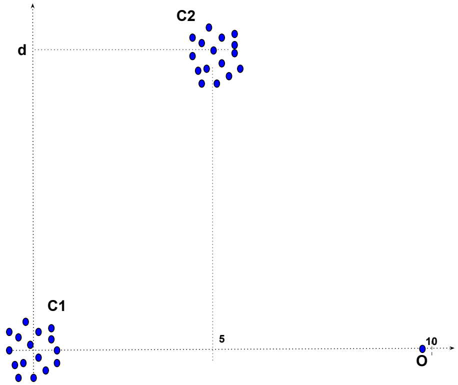
$$\begin{aligned}\text{Expected SSE } \approx & \left(\frac{1}{n} + \frac{d^2}{\Phi_X}\right) \cdot \left(\frac{n}{2} + d^2 \cdot \frac{n}{2}\right) + \\ & \frac{1}{\Phi_X} \cdot \left(\frac{n}{2} + 100 \cdot \frac{n}{2} + d^2\right) + \\ & \left(1 - \frac{1}{n} - \frac{d^2}{\Phi_X} - \frac{1}{\Phi_X}\right) \cdot \left(\frac{n}{2} + \frac{n}{2} + d^2\right)\end{aligned}$$

We obtain:

$$\text{Expected SSE } \approx c_1 \cdot d^2 + c_2 + n + d^2 + c_3 \cdot d^2 \approx n + d^2,$$

where c_1, c_2, c_3 are some constants. As a result, when n is large k-means++ performs much better than the deterministic algorithm (whose SSE$\approx \frac{n}{2} \cdot d^2$).

Input Points: case 2



In this case, we can use a similar argument and prove similar bounds for SSE of k-means++. Note that in this case the deterministic algorithm might perform slightly better.

k-means++: conclusions

- k-means++ is able to distinguish between a single outlier and a whole cluster of points. This is not the case for the deterministic algorithm (which fails in the first case).
- k-means++ performs well on average, that is, several runs of the algorithm might be necessary so as to obtain a good SSE error.