COMP229: Introduction to Data Science
Lecture 26: centroid-based clustering

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

Approaches to clustering

Many clustering algorithms split into classes.

- *Hierarchical* clustering outputs a hierarchy of clusters, e.g. the single-linkage clustering.
- *Centroid-based* clustering optimises centres of clusters, e.g. we'll discuss k-means.
- *Distribution-based* clustering, e.g. the expectation-maximisation algorithm for Gaussian (normal) mixture models.
- *Density-based* clustering defines clusters as areas of higher density, others are outliers.

The k-means clustering objective

Definition 26.1. For a cloud (cluster) of points $C = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$, its *centre* is $\bar{C} = \dfrac{1}{n}\sum_{i=1}^{n} p_i$, where each $p_i \in \mathbb{R}^d$ is considered as a vector.

The *k-means clustering* aims to split a cloud C into disjoint clusters $C_1, \ldots, C_k$ to minimise $\sum_{i=1}^{k} \sum_{p \in C_i} d^2(p, \bar{C}_i)$, d is the Euclidean distance.

Find the centre of the cloud: $(3, 2), (-4, -1), (1, -5),$ $(-1, -4), (2, -3), (4, 1), (-5, 4), (-3, 5), (5, -2), (-2, 3)$.

The neighbourhood cluster of a centre

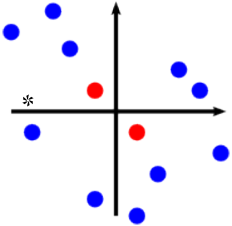
The centre of the above cloud in $\mathbb{R}^2$ is $(0, 0)$.

Definition 26.2. For a cloud $C = \{p_1, \dots, p_n\}$ in $\mathbb{R}^d$ and some given centres $q_1, \dots, q_k \in \mathbb{R}^d$, the *neighbourhood* of a point q_i within the cloud C is $N(q_i) = \{p \in C \, : \, d(p, q_i) \leq d(p, q_j) \text{ for } j \neq i\}$.

If a point $p \in C$ is the mid-point between q_i, q_j, one can make a random (or other) choice for p.

Find the neighbourhoods of the centres $(1, -1)$ and $(-1, 1)$ in the cloud: $(3, 2)$, $(-4, -1)$, $(1, -5)$, $(-1, -4), (2, -3), (4, 1), (-5, 4), (-3, 5), (5, -2), (-2, 3)$.

Initialisation of k-means clustering



The neighbourhood of the centre $(-1, 1)$ consists of $(-4, -1), (-2, 3), (-5, 4), (-3, 5)$, while $(1, -1)$ attracts all others.

Input: a cloud $C \subset \mathbb{R}^d$, a number of k of clusters.

Initialisation of centres. The *Forgy* method chooses k initial centres as random points of C (usually spread out). The *Random Partition* method randomly splits a cloud into k subsets and takes their centres (often close to the centre of C).

Standard Lloyd's heuristic algorithm

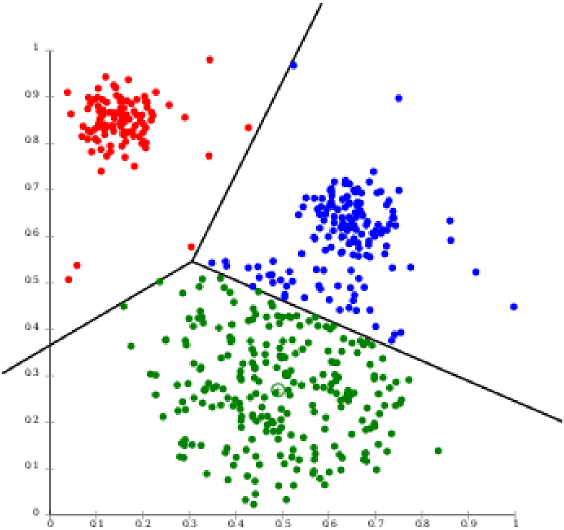
Alternate between the two steps below.

1) **Assignment**: for current centres $q_1, \dots, q_k$, split the cloud C into their neighbourhoods $N(q_1), \dots$

2) **Update**: for any new cluster C_j, its centre is updated as the mean point: $\bar{C}_j = \frac{1}{n} \sum\limits_{p \in C_j} p$.

Above Lloyd's algorithm has converged when the assignments no longer change. There is no guarantee that an optimal partition is found.

k-means partitions the data space



The assignment step of Lloyd's algorithm splits the data cloud into in neigh-bourhoods of cluster centres (three subclouds in the picture).

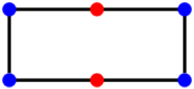
Complexity of k-means algorithms

For a fixed dimension d and a number k of clusters, an optimal partition can be found in time $O(n^{dk+1})$, i.e. the number of operations is proportional to n^{dk+1} when $n \to +\infty$.

Lloyd's algorithm has time $O(nkdi)$, where i is the number of iterations. In the worst-case $i = 2^{\Omega(n)}$, where $\Omega(n)$ denotes a function that grows proportional to n or faster. In practice i is often small on data split into well-separated groups.

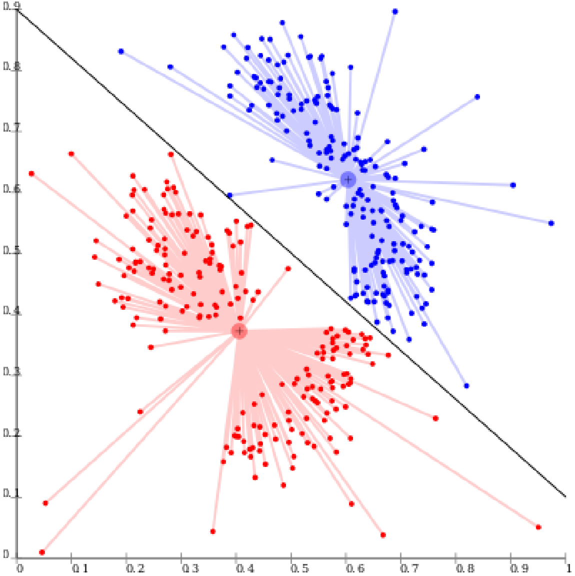
Arbitrarily bad k-means clustering

Different initial centres may lead to different outputs, so k-means is often run with different initialisations. Let the cloud C be 4 vertices of an axes-aligned rectangle. If 2 initial centres are the mid-points of the horizontal edges, then 2-means outputs these two centres without iterations.



If the horizontal sides are much longer than the vertical ones, this output is far away from the optimal clustering with centres at the mid-points of the vertical edges.

k-means requires a good value of k



The key drawback of k-means: we need a good value of k. The cloud in the picture has 3 high-density regions, but 2-means clustering outputs only 2 bad clusters if $k = 2$ is chosen.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Is the running time of k-means clustering polynomial in the number n of points?

Answer to the quiz and summary

Answer. $O(n^{dk+1})$ is polynomial for fixed d, k.

Here are the steps of Lloyd's heuristic algorithm.

- Initialise k centres of clusters in a cloud C.
- To each centre assign all points of C that are closer to this centre than to all others.
- Re-compute the centre of every cluster that was updated above and re-assign all points.
- Stop when centres of clusters don't change or a maximum number of iterations is reached.