As far as I know, the "mean" of a cluster and the centroid of a single cluster are the same thing, though the term "centroid" might be a little more precise than "mean" when dealing with multivariate data.

To find the centroid, one computes the (arithmetic) mean of the points' positions separately for each dimension. For example, if you had points at:

- (-1, 10, 3),
- (0, 5, 2), and
- (1, 20, 10),

then the centroid would be located at ((-1+0+1)/3, (10+5+20)/3, (3+2+10)/3), which simplifies (0, 11 2/3, 5). (NB: The centroid does not have to be--and rarely is---one of the original data points)

The centroid is also sometimes called the center of mass or barycenter, based on its physical interpretation (it's the center of mass of an object defined by the points). Like the mean, the centroid's location minimizes the sum-squared distance from the other points.

A related idea is the **medoid**, which is the data point that is "least dissimilar" from all of the other data points. Unlike the centroid, the medoid has to be one of the original points. You may also be interested in the **geometric median** which is analgous to the median, but for multivariate data. These are both different from the centroid.

However, as Gabe points out in his answer, there is a difference between the "centroid distance" and the "average distance" when you're comparing clusters. The **centroid distance** between cluster A and B is simply the distance between $\text{centroid}(A)$ and $\text{centroid}(B)$. The **average distance** is calculated by finding the average pairwise distance between the points in each cluster. In other words, for every point a_i in cluster A, you calculate $\text{dist}(a_i, b_1)$, $\text{dist}(a_i, b_2)$, ... $\text{dist}(a_i, b_n)$ and average them all together.

edited Apr 13 '17 at 12:44
Community ♦
1

answered Mar 9 '13 at 1:40
Matt Krause
13.3k 2 34 75

Under what conditions the centroid and the medoid be identical? And also why the centroid is a good representative of a set of points? – dkr Jan 21 at 20:57

@dkr, You might want to ask this as a new question to get more (and more in-depth) responses. That said, the difference boils down to two things: 1) the thing to be minimized (squared distance/L2 norm for the centroid, absolute distance/L1 norm for mediod) and 2) Whether the output can be any point (centroid) or must be in the data set (mediod). You can imagine cases where they'll be the same, but in general, they will not. The centroid is "good" for the same reasons the mean is (smallest sum-squared distance to the points) and also has similar drawbacks (e.g., not robust against outliers).
– Matt Krause Jan 21 at 23:00