Computer Science 134C
Introduction to Computer Science, in Python
Lecture #18 (Classes III)
October 26

Keywords
clustering, container class, k-means, randomized algorithm

We continue building classes.

1. Questions?

2. Today, a more complicated class for identifying and keeping track of *clusters* of related data.

3. We'll use the *k-means* algorithm for grouping data into k clusters, with immutable data values "close to" around "means". Here's the outline of the approach:

    (a) Guess or pick k values to be the respective *representative* values of k groups or *clusters* of your data. It is unlikely that these k values are *actually* good representative values of your data.

    (b) Now *classify* your data points: find the representative value they're closest to and place them in that representative's cluster.

    (c) While the k representatives are close to all the values in their respective clusters, there are probably better representative values. Compute the k mean values of the clusters and use these as the new representatives.

    (d) Recluster the data based on these new k mean values.

    (e) Repeat until data stops moving around, or variance is reduced, or simply a pre-determined number of times.

4. Subtle points:

    (a) We'll want to think about how a class might be used to help with this process.

    (b) Since clustering depends on the relationships between data points it will be important to make sure that the user cannot change the data once it has been clustered.

    (c) We should provide ready access to the k means.

    (d) We should, given a mean, be able to access its cluster elements.

    (e) It would be nice to be able to classify new points, on-the-fly.

⋆