Introduction to Machine Learning Program Assignment #2

TA's name: 林裕庭 Deadline: 2018/11/09 (Fri) 23:59:59

TA's email: kartglin.iie07g@nctu.edu.tw

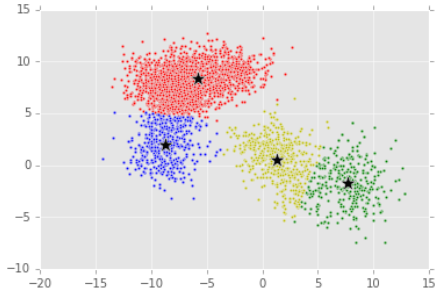
This program assignment aims to help you understand the K-means and Kd-tree implementation.

I. K-means Problem

You will get a dataset (**data_noah.csv**). It is Noah Syndergaard's pitches that have been tracked by the PITCHf/x system in the MLB Regular Season.

You have to do the following:

1. Dataset including 1322 number of instances with many attributes.
2. ***Don't*** **use the library related to K-means.**
   (i.e. Construct a K-means function by yourself).
3. Use **Attribute x** (horizontal movement) and **y** (vertical movement) to partition 1322 pitches into 3 clusters.
4. 3 clusters will represent FF (four-seam fastball), CH (changeup) and CU (curveball).
5. **Construct a cost function to check the accuracy of pitch types.**
6. **Generate a figure** to show the result of K-Means clustering.
   For example:



7. Try to use another two or more attributes (like speed) to partition.
   Don't worry whether the accuracy is high or not!
8. Try to explain why k = 3 is the best, and write in your report.
9. Show your **code**, **accuracy**, the reason of k = 3 and the result of K-Means clustering **(figure) in your report**.

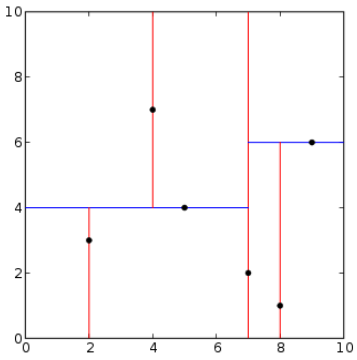
- If you are interested, you can get more information of pitches from ***brooksbaseball.***
  (http://www.brooksbaseball.net/landing.php?player=592789)

II. Kd-tree Problem

You will get a set of points (**points.txt**) in the unit square (all points have x-coordinates and y-coordinates). You have to build a 2d-tree.

You have to do the following:

1. **You *can* use the library related to Kd-tree.**
2. **Draw a 2d-tree divides the unit square (Use two colors).**
   For example:



3. Show your **code** and the result of 2d-tree **(figure) in your report**.

- If you are interested, you can construct a Kd-tree function by yourself.
  - Calculate the variance of this two dimensions and select **the big one** as axis-aligned splitting planes.
  - Then, sort points in the given set and choose **median** as pivot element where you should split.
  - As one moves down the tree, one cycles through the axes used to select the splitting planes. (For example, in a 2-dimensional tree, the root would have an x-aligned plane, the root's children would have y-aligned planes, the root's grandchildren would have x-aligned planes, and so on.)

III. Report & Scoring

This is a team-based program assignment, so **one team should only submit one report and one source code to E3**.

The report should contain the following:

1. What environments the members are using (5%)
2. K-means code (30%)
3. Cost function and accuracy (15%)
4. The result of K-Means clustering (10%)
5. Use another two or more attributes to partition and the reason of k = 3 (10%)
6. Kd-tree code (15%)
7. The result of Kd-tree (15%)

There are some rules to follow:

1. C / C++ / Java / Python / Matlab are allowed to use. For visualization, Excel or other programs are allowed.
2. Report format should be **PDF**.
3. **Attach your code when you are submitting.**
4. No cheating and plagiarizing.
5. **Delay：Your score *= 0.8**