# CS618: Assignment 5

Total Marks: 100

Due on: 11th March, 2015, 01:00am

This assignment is to help understand the basics of distance-based indexing using *VP-trees* and *GH-trees*.

Implement a basic *VP-tree*.
Choose the root of each subtree as the one with the largest variance in distance from a *sample* of objects. Assume the structure to be in memory.

Enable it to handle range queries and kNN queries. (Insertions and deletions may be ignored.)

Instead of VH-tree, you may complete the exercise for a basic *GH-tree*, or you may do it additionally.
Choose the two pivots of each subtree as the pair with the largest distance between them from a *sample* of objects. Again, assume the structure to be in memory.

Use the file `assgn5_data.txt` to inject the points. It contains $10^6$ 2-dimensional points. Use the $L_2$ norm as the distance between the points.

*Never* use the coordinates of the points.

Use the file `assgn5_querysample.txt` to read the queries. The queries use the same distance function $L_2$. The queries have the following formats:

| Operation | Code | Details | |
|---|---|---|---|
| Range query | 2 | Query object | Range |
| kNN query | 3 | Query object | Number of nearest neighbors |

Enable the program to output timing results string from the reading of a query to solving it. Do *not* include the time to print it.

Compare the two structures if you have implemented both.
Report the following times for both the structures and for each type of operation: (i) minimum, (ii) maximum, (iii) average, (iv) standard deviation.
Report also the number of distance computations times for both the structures and for each type of operation: (i) minimum, (ii) maximum, (iii) average, (iv) standard deviation.

Repeat the entire set of exercises for both the structures by using the *Mahalanobis distance*. The Mahalanobis distance between two points $\vec{x}$ and $\vec{y}$ is defined as $M(x, y) = \sqrt{(\vec{x} - \vec{y})A(\vec{x} - \vec{y})^T}$ where $A$ is the inverse of the correlation matrix between the dimensions in the dataset. For your convenience, the matrix $A$ is given in `assgn5_matrix.txt`. The points $\vec{x}$ and $\vec{y}$ are assumed to be in row-format here.
As an example, if $\vec{x} = (0.2, 0.4); \vec{y} = (0.6, 0.3); A = [2\ 0.8;\ 0.8\ 2]$,
then $M(x, y) = \sqrt{[0.4\ 0.1][2\ 0.8;\ 0.8\ 2][0.4\ 0.1]^T} = \sqrt{0.4040} = 0.6356$.

What do you conclude?

Submit the program and the answers through the submission portal only. You must name your submission `studentno_assgn5.zip`. The student numbers (which are *not* the roll numbers) are 2-digit codes and are available from the course website.

We will evaluate the program by running a query file with the same format (and the same matrix file) as the sample one. Marks will be deducted for wrong answers.