# Big Data Analytics

**Homework 3 (X-Tree, M-Tree, and OMNI-Family)**

In this homework, there are 4 questions + 1 bonus question, covering the X-tree, M-tree, and OMNI-family. If you can answer the bonus question correctly, you can obtain 20 extra points. The maximum mark for this homework is **120 points**, which will be later scaled.

1. Please list two major differences (or improvement) of X-tree from R\*-tree. [10 points]

**One of the biggest advantages of a X-tree is its ability to handle high dimensionality of data. Unlike the R-tree family whereas dimensionality increases and performance decreases, the X-tree can handle and perform with dimensionality on a much higher and more productive scale. Overlapping data also presents a problem for R-Tree family. The overlapping data in an X-tree is more geared towards avoiding the overlapping without allowing the tree to degenerate, using super nodes to manage the data at a more effective and precise way. Overall, the X-Tree has more adaptability towards dimensionality, overlap and natural abundance of data similar in structure.**

2. Please list at least 2 distance functions that are ***metric distances***, and at least 2 distance functions (or similarity measures) that are ***non-metric distances***. For each distance function, give the reason why it is (or is not) a metric distance, and give references or URL links. (***Hint:*** *please search on the Web or Wikipedia to find the answers*) [20 points].

**Two Metric Distances:**

**1. Euclidean Distance: Uses metric distance to calculate distance between two points in a plane or a cluster of data. Is a branch of the Minkowski distance.**

**h**[**ttps://towardsdatascience.com/importance-of-distance-metrics-in-m**](ttps://towardsdatascience.com/importance-of-distance-metrics-in-m)**achine-learning-modelling-e51395ffe60d**

**2. Manhattan Distance: Uses metric distancing to calculate distance between two data points in a grid like path. Is a branch of the Minkowski distance.**

**h**[**ttps://towardsdatascience.com/importance-of-distance-metrics-in-m**](ttps://towardsdatascience.com/importance-of-distance-metrics-in-m)**achine-learning-modelling-e51395ffe60d**

**Two Non-Metric Distances:**

**1. Fisher Criterion: This is non-metric because all distances are not equal if parts where in metric distances this needs to be the case. It is used to find the measure of suitability between two points in a single variable.**

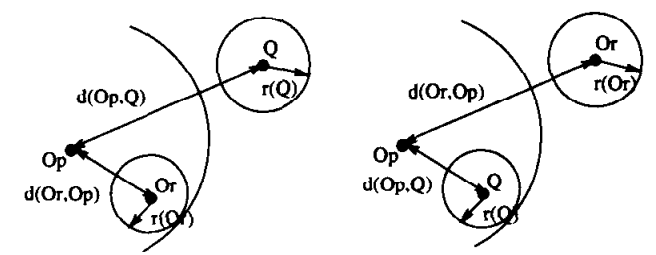
[**http://37steps.com/2202/non-metric-disreps/**](http://37steps.com/2202/non-metric-disreps/)

**2. Image Retrievals: Image retrievals are non metric because of the processing of physical properties not on a vector or in a cluster. The differences in say three different images violate the triangle inequality.**

[**https://www.cs.umd.edu/~djacobs/pubs\_files/non-metric.pdf**](https://www.cs.umd.edu/~djacobs/pubs_files/non-metric.pdf)

3. Please read the lecture slide of Chapter 5, "Range Queries Over M-Tree", and prove the pruning strategy for the range query below (***Hint****: use the triangle inequality*) [30 points]:

If |*d*(*Op*, *Q*)-*d*(*Or*, *Op*)|>*r*(*Q*)+*r*(*Or*), then *d*(*Or*, *Q*) > *r*(*Q*) + *r*(*Or*) holds and node centered at *Or* with radius *r*(*Or*)can be safely pruned.



**Triangle Inequality = dist(x,y) + dist(y,z) ≥ dist(x,z)**

**X=Qp Y=Qr Z=Q examples numbers Qp=4 Qr=8 Q=6**

**Dist(Qp,Qr)+Dist(Qr,Q) ≥ Dist(Qp,Q)**

**Dist(4,6)+Dist(8,6) ≥ Dist(4,6)**

4. **(The Curse of Dimensionality)** [40 points]

4a. What is the curse of the dimensionality? Please provide the reason for the dimensionality curse. [10 points]

**The curse of dimensionality is the phenomena that happens when analyzing and organizing data in high-dimensional spaces that do not occur in low dimensional settings. The data has to many features to process. The reason for the curse is the need and desire to run functions over data that has so many features(information) that is just is not possible to do in one simple function and algorithm. A reason is the wants and needs of stakeholders who have money and time invested wanting results quick on data that has to many features resulting in a dimensionality curse when looking for conclusions.**

[**https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e**](https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e)

[**https://kent.ares.atlas-sys.com/areslms//ares.dll?SessionID=B052110985H&Action=10&Type=10&Value=160499**](https://kent.ares.atlas-sys.com/areslms//ares.dll?SessionID=B052110985H&Action=10&Type=10&Value=160499)

4b. Read Section 2 of the following paper, and write a short survey about existing dimensionality reduction techniques and high dimensional data structures mentioned in this section (*Please cite reference papers in your survey and provide a list of reference papers after the survey. You may need to read abstract or introduction of some reference papers, if they are unclear in the section.* ***Note:*** *please use your own words to describe the techniques;* ***DO NOT*** *copy any sentences from the paper*). [30 points]

H. T. Shen, X. Zhou, and A. Zhou. An adaptive and dynamic dimensionality reduction method for high-dimensional indexing. In *VLDBJ*, 2006. *Located in the Library Course Reserves on the left-hand course menu.*

**Global Dimensionality Reduction consist of reducing dimensionality of data and then indexing the reduced dimensionality data into separate structures (Chakrabarti and Mehrotra, 2000). Reducing the data, you are clustering data into locally correlated data in a new structure of the algorithm to search. The ability of the algorithm now can shoot back false positives not returned to the user so the end user does not have to worry about data coming back that is not correlated in dimensionality (Chakrabarti and Mehrotra, 2000). Local Dimensionality Reduction is a step within the GDR as it searches for locally already correlated data, began to be mentioned above (Chakrabarti and Mehrotra, 2000). Performing the reduction and clustering on smaller, each cluster/node is where LDR gets its name, a more precise algorithm for wanting to know more exacts rather than sums (Chakrabarti and Mehrotra, 2000). Both GDR and LDR are effective in capturing dimensionality and turning it into a positive and smart algorithm with precise results while staying away from the dimensionality curse.**

**High-dimensional structures (H.T. Shen, et.al., 2006) is broken down into data approximation and data transformation (H.T. Shen, et.al., 2006) which are two common approaches of high-dimensional reduction. Data approximation is using smaller data points to help narrow down and calculate better representations of the data (H.T. Shen, et.al., 2006). VA-File is a approach off of data approximation. Vector approximation file helps to approximate data leading to an improved performance as dimensionality increases in the data (Weber, et. al., 1998). VA-File reduces the amount of data that needs to be read during searches, that is how the performance increases in this case (Weber, et. al., 1998). In higher dimensionality, as X-Tree trumps R-Trees, this technique of VA-File trumps X-Trees in its performance with high-dimensionality (Weber, et. al., 1998). For data transformation, this helps with converting data from one format or structure into one more useable for the end goals (Wikipedia). A popular data transformation technique is the Pyramid technique. The pyramid technique splits data into two-dimensional pyramids then further breaks down the pyramid to transform data into proper fitting groups and clusters (Berchtold, et.al., 1998). Dimensionality does become a problem further into this technique so newly developed techniques of PROCLUS, OptGrid and Wavlet transformations have been developed to deal with larger, more dimensional data sets (H.T. Shen, et.al., 2006).**

**Berchtold, S., Böhm, C., Kriegel, H.-P.: The pyramid-technique: towards breaking the curse of dimensionality. In: SIGMOD, pp. 142–153 (1998)**

**Chakrabarti, K., Mehrotra, S.: Local dimensionality reduction: a new approach to indexing high dimensional spaces. In: VLDB, pp. 89–100 (2000)**

**H. T. Shen, X. Zhou, and A. Zhou. An adaptive and dynamic dimensionality reduction method for high-dimensional indexing (2006)**

**Weber, R., Schek, H., Blott, S.: A quantitative analysis and per- formance study for similarity search methods in high dimensional spaces. In: VLDB, pp. 194–205 (1998)**

**Wikipedia, https://en.wikipedia.org/wiki/Data\_transformation**

**Bonus Question [20 extra points]**

5. Read Section 2.2 of the following paper, as well as the cited papers in this subsection, and write a short survey about the *intrinsic dimensionality*. (***Note:*** *please use your own words to describe the problem definition and solutions;* ***DO NOT*** *copy any sentences from the paper*).

R.F.S. Filho, A. Traina, C. Traina, and C. Faloutsos. Similarity search without tears: the OMNI-family of all-purpose access methods. In *ICDE*, 2001. *Located in the Library Course Reserves on the left-hand course menu.*

**Intrinsic dimensionality gives precision in the selectivity estimation for NN queries (R.F.S. et. al., 2001). Variables need to be in minimal representation of the data (Wikipedia). An example is when in an equation, a variable can be described as being the single variable of its kind in the equation, it will be intrinsic (Wikipedia) and example is f(x^1+x^2/4), here you have the f alone, an intrinsic dimension being created when solved for the f. Another idea of intrinsic dimensionality is to extract data from local clusters/groupings early on in the data working (Wikström, et. al., 2014). Time is money in todays business world, so by getting the intrinsic dimensionality sought, found, and used, is key to producing more accurate data and data in a more timely fashion. Local Intrinsic Dimensionality is when one isolated variable is close to 0 within its local low dimensional cluster/group. The importance of intrinsic dimensionality can be a stepping stone of giving good approximation to begin to tell the story of the data before it gets further worked on and further more stories begin told from the data set.**

**R.F.S. Filho, A. Traina, C. Traina, and C. Faloutsos. Similarity search without tears: the OMNI-family of all-purpose access methods. In *ICDE*, 2001**

**Wikipedia,** [**https://en.wikipedia.org/wiki/Intrinsic\_dimension**](https://en.wikipedia.org/wiki/Intrinsic_dimension)

**Wikström, J. Georgoulas, G. Thucydides , M. Seferiadis, A., Intelligent data analysis of instrumented gait data in stroke patients—A systematic review, Computers in Biology and Medicine, Volume 51, 2014,Pages 61-72**

## Submitting Your Assignment

*All work must be your own. Copying other people’s work or from the Internet is a form of plagiarism and will be prosecuted as such.*

You may submit a Microsoft Word (.docx) document as an attachment. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

You can submit multiple times and only the last submission attempt will be considered for grading.

* Submissions sent by email will NOT be accepted.