Quiz 5 Unsupervised Learning

1. Summarize a use case where you must use Manhattan Distance as the similarity measure in unsupervised learning.

*The use of Manhattan distance depends a lot on the kind of co-ordinate system that your dataset is using. While Euclidean distance gives the shortest or minimum distance between two points, Manhattan has specific implementations.*

*For example, if we were to use a Chess dataset, the use of Manhattan distance is more appropriate than Euclidean distance. Another use would be when are interested in knowing the distance between houses which are few blocks apart.*

*Also, you might want to consider Manhattan distance if the input variables are not similar in type (such as age, gender, height, etc.). Due to the curse of dimensionality, we know that Euclidean distance becomes a poor choice as the number of dimensions increases.*

*So, in a nutshell: Manhattan distance generally works only if the points are arranged in the form of a grid and the problem which we are working on gives more priority to the distance between the points only along with the grids, but not the geometric distance.*

<https://datascience.stackexchange.com/questions/20075/when-would-one-use-manhattan-distance-as-opposed-to-euclidean-distance>

*According to this interesting paper, Manhattan distance (L1 norm) may be preferable to Euclidean distance (L2 norm) for the case of high dimensional data.*

*The authors of the paper even go a step further and suggest using Lk norm distances, with a fractional value of k, for very high dimensional data in order to improve the results of distance-based algorithms, like clustering.*

<https://bib.dbvis.de/uploadedFiles/155.pdf>

1. Summarize (1 paragraph each) any 5 unsupervised learning algorithms, how they work and what they are used for.
   1. K-Means
   2. Hierarchical clustering
   3. PCA
   4. Frequent Item
   5. Singular Value Decomposition

<https://www.guru99.com/unsupervised-machine-learning.html>

This article provides really good examples!

<https://theappsolutions.com/blog/development/unsupervised-machine-learning/>

1. Construct a simple X-Y data set from this data set using X = Age and Y = BMI

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| number | age | sex | bmi | children | smoker | region | charges |
| 1 | 19 | female | 27.9 | 0 | yes | southwest | 16884.92 |
| 2 | 18 | male | 33.77 | 1 | no | southeast | 1725.552 |
| 3 | 28 | male | 33 | 3 | no | southeast | 4449.462 |
| 4 | 33 | male | 22.705 | 0 | no | northwest | 21984.47 |
| 5 | 32 | male | 28.88 | 0 | no | northwest | 3866.855 |
| 6 | 31 | female | 25.74 | 0 | no | southeast | 3756.622 |
| 7 | 46 | female | 33.44 | 1 | no | southeast | 8240.59 |
| 8 | 37 | female | 27.74 | 3 | no | northwest | 7281.506 |
| 9 | 37 | male | 29.83 | 2 | no | northeast | 6406.411 |
| 10 | 60 | female | 25.84 | 0 | no | northwest | 28923.14 |
| 11 | 25 | male | 26.22 | 0 | no | northeast | 2721.321 |
| 12 | 62 | female | 26.29 | 0 | yes | southeast | 27808.73 |
| 13 | 23 | male | 34.4 | 0 | no | southwest | 1826.843 |
| 14 | 56 | female | 39.82 | 0 | no | southeast | 11090.72 |
| 15 | 27 | male | 42.13 | 0 | yes | southeast | 39611.76 |
| 16 | 19 | male | 24.6 | 1 | no | southwest | 1837.237 |
| 17 | 52 | female | 30.78 | 1 | no | northeast | 10797.34 |
| 18 | 23 | male | 23.845 | 0 | no | northeast | 2395.172 |
| 19 | 56 | male | 40.3 | 0 | no | southwest | 10602.39 |
| 20 | 30 | male | 35.3 | 0 | yes | southwest | 36837.47 |
| 21 | 60 | female | 36.005 | 0 | no | northeast | 13228.85 |
| 22 | 30 | female | 32.4 | 1 | no | southwest | 4149.736 |
| 23 | 18 | male | 34.1 | 0 | no | southeast | 1137.011 |
| 24 | 34 | female | 31.92 | 1 | yes | northeast | 37701.88 |
| 25 | 37 | male | 28.025 | 2 | no | northwest | 6203.902 |
| 26 | 59 | female | 27.72 | 3 | no | southeast | 14001.13 |
| 27 | 63 | female | 23.085 | 0 | no | northeast | 14451.84 |
| 28 | 55 | female | 32.775 | 2 | no | northwest | 12268.63 |
| 29 | 23 | male | 17.385 | 1 | no | northwest | 2775.192 |
| 30 | 31 | male | 36.3 | 2 | yes | southwest | 38711 |

Make number of rows = 20.

Explain how you would construct a dendrogram using hierarchical clustering.

See <https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8>