Data Mining I: Homework 2
Similarity Measures on time Series and Graphs

Toby Law

20 October 2019

Exercise 1.c

Compare and discuss the results of the `DTW` and `Manhattan` distances on separating abnormal and normal heartbeats.

Both `DTW` and `Manhattan` distances are able to report a greater average distance for normal : abnormal comparisons when compared to comparisons within the group of normal heartbeats. However, we don't see a similarly significant difference in distance between abnormal : abnormal and abnormal : normal comparisons. The average distance of abnormal : normal comparisons is not larger than comparisons within the group of abnormal heartbeats, not until `DTW` distance ($w \geq 25$). This might be because there is inherently greater fluctuation for the time series measurements of abnormal heartbeats, since they represent the condition of a diseased patient.

Exercise 1.d

Discuss the effect that hyperparameter w has on the `DTW` distance and its ability to separate abnormal and normal heartbeats.

The hyperparameter w constricts the distance between the time points in the two time series that can be aligned together. The greater w is, the greater the degrees of freedom in the alignment, and when w reaches infinity, we are essentially finding the minimal distance between any two time points of the two time series, and aligning them together, without considering how far apart they are in time. Therefore, we can see that as w increases, `DTW` distance decreases for all comparisons, and the distance contrast between types of comparisons also decrease.

Exercise 1.e

Is the `DTW` distance a metric? If not, give examples showing which conditions are not satisfied.

$$DTW_d(t_1, t_2) = \min_{w \in W(m,n)} \sum_{l=1}^{L(w)} d(t_1[w_l^{(1)}], t_2[w_l^{(2)}])$$

$$Let \quad w_l = (w_l^{(1)}, w_l^{(2)}) = (i, j).$$

$$d(t_1, t_2) = (t_1[i], t_2[j]) = |t_1[i] - t_2[j]| + min \begin{cases} |t_1[i-1] - t_2[j-1]| \\ |t_1[i] - t_2[j-1]| \\ |t_1[i-1] t_2[j]| \end{cases}$$

Let t_1 be (1, 2, 3), t_2 be (4, 5, 6) and t_3 be (7, 8, 9).

$$\begin{aligned} DTW_d(t_1, t_3) &= 6 + (6 + 5) + (6 + 5) = 28 \\ DTW_d(t_1, t_2) &= 3 + (3 + 2) + (3 + 2) = 13 \\ DTW_d(t_2, t_3) &= 3 + (3 + 2) + (3 + 2) = 13 \\ DTW_d(t_1, t_3) &> DTW_d(t_1, t_2) + DTW_d(t_2, t_3) \end{aligned}$$

DTW does not satisfy the triangle inequality, therefore, it is not a metric.

Exercise 1.f

What is the runtime complexity of computing the `DTW` distance with w-constrained warping? You might consider that $m = n$.

To compute the values of matrix C, the runtime complexity is equal to $O(N^2)$. With w-constrained warping, for each step we have to execute an extra computation to determine if $|i - j| < w$, which makes the runtime complexity $O(N^3)$.

Exercise 2.c

What is the runtime complexity of Floyd-Warshall's algorithm? What is the runtime complexity of `SPKernel`?

The runtime complexity of Floyd-Warshall's algorithm is $O(N^3)$. There are N^2 possible connections between nodes i and j, and we need to compute the path between node i and j through all possible intermediate nodes k to find the shortest path for each of them.
The runtime complexity of `SPKernel` is $O(N^4)$. There are N^2 edges in each shortest path matrix for each graph, and we need to compare $N^2 \times N^2$ edges to obtain the `SPKernel`.