MM811 Winter 2016 Assignment 5 Report

Name: Xinyao Sun

Experiment Environment:

I implemented each task through Matlab script and running in Matlab 2015b for the experiment. The experiment machine: CPU: i7 dual cores 2.5Ghz, RAM: 32GB, System: 64bit Windows 10: workload: 15% RAM and 5% CPU usage.

Task 1:

(a) Accuracy by using 1-NN with **full** Euclidean distance:

Dataset	Accuracy (%)
ECG	92.49
NonInvasiveFatalECG_Thorax1	82.90

(b) Cost time measurement for computing one pair of sequence's **full** Euclidean distance with the mean and standard deviations:

Dataset	Mean Cost time (ms)	Std for Cost time (ms)
ECG	0.002913	0.001900
NonInvasiveFatalECG_Thorax1	0.011579	0.005777

Task 2:

(a) Accuracy by using 1-NN with **early abandoning** Euclidean distance:

Dataset	Accuracy (%)
ECG	92.49
NonInvasiveFatalECG_Thorax1	82.90

(b) Cost time measurement for computing one pair of sequence's **early abandoning** Euclidean distance with the mean and standard deviations:

Dataset	Mean Cost time (ms)	Std for Cost time (ms)
ECG	0.001838	0.003774
NonInvasiveFatalECG_Thorax1	0.001661	0.002558

Discussion: Using the early abandoning strategy reduced the calculation time and kept the accuracy, especially for long length sequence.

Task 3:

(a) Validation for DFT function, where I used *fft* Matlab function for DFT computation for each sequence during this Task. To create five pairs validation sequence, I chosen first sequence from ECG_TEST paired with first five sequences from ECG_TRAIN. Moreover, because original sequence is too long, therefore, I just used first four entries.

Six *DFT* sequence:

Sequence ID:	***DFT* sequence**
1	[-1.8529 + 0.0000i, 5.8049 - 4.8524i, 5.0063 + 0.0000i, 5.8049 + 4.8524i]
2	[-11.0634 + 0.0000i 3.6614 - 1.5226i 3.2906 + 0.0000i 3.6614 + 1.5226i]
3	[-13.8901 + 0.0000i 3.1849 - 0.5098i 3.1167 + 0.0000i 3.1849 + 0.5098i]
4	[-11.6189 + 0.0000i 3.3071 - 1.9906i 2.7363 + 0.0000i 3.3071 + 1.9906i]
5	[-9.3591 + 0.0000i 4.1069 - 2.4044i 3.1073 + 0.0000i 4.1069 + 2.4044i]
6	[-6.4321 + 0.0000i 3.1850 - 3.0991i 3.2630 + 0.0000i 3.1850 + 3.0991i]

Respective original sequence:

Sequence ID:	***Original* sequence**
1	[3.6908, 0.7114, -2.1141, -4.1410]
2	[-0.1125, -2.8272, -3.7739, -4.3498]
3	[-1.1009, -3.9968, -4.2858, -4.5066]
4	[-0.5671, -2.5935, -3.8742, -4.5841]
5	[0.4905, -1.9144, -3.6164, -4.3188]
6	[0.8002, -0.8742, -2.3848, -3.9733]

Five pairs with sequence ID and its Euclidean distance:

Pairs:	***DFT* sequence Distance**	***Original* sequence Distance**
(1,2)	10.9152	5.4576
(1,3)	14.1390	7.0695
(1,4)	11.3748	5.6874
(1,5)	8.8149	4.4074
(1,6)	6.66245	3.3123

Discussion: The order of the sequence 2-6 from closest to sequence 1 to farthest to sequence 1, stay same between using the ***original*** sequence and using their respective *DFT* sequence.

(b) Accuracy by using 1-NN with **early abandoning** Euclidean distance on *DFT* sequence:

Dataset	**Accuracy (%)**
ECG	92.49
NonInvasiveFatalECG_Thorax1	82.90

Cost time measurement for computing one pair of sequence's **early abandoning** Euclidean distance with the mean and standard deviations:

Dataset	**Mean Cost time (ms)**	**Std for Cost time (ms)**
ECG	0.004117	0.003802
NonInvasiveFatalECG_Thorax1	0.004877	0.005411

Discussion: Comparing the table in Task 2, the mean cost time spent for one pair of sequence becomes larger after using the *DFT* sequence. I think the reason is calculating the Euclidean distance is more expensive for complex number vectors than in real numbers.

(c) Accuracy and time measurement by using 1-NN with **full** Euclidean distance on ***DFT*** sequence with using different number of coefficients:

ECG Dataset:

Num. of coefficients	**Accuracy (%)**	**Mean Cost time (ms)**	**Std for Cost time (ms)**
5	91.71	0.004213	0.002987
10	92.31	0.004934	0.003084
20	92.31	0.006485	0.004392

NonInvasiveFatalECG_Thorax1 Dataset:

Num. of coefficients	**Accuracy (%)**	**Mean Cost time (ms)**	**Std for Cost time (ms)**
5	73.99	0.005049	0.002780
10	81.32	0.005780	0.003091
20	82.90	0.007303	0.003632

Discussion: While using the **full** Euclidean distance on DFT sequence, as the number of used coefficients increasing the accuracy has been increased, at the same time the mean cost time for each pair's distance computing also been increased. Moreover, for using 10 and 20 first DFT coefficients, the accuracy has extremely close to the accuracy using full sequence. Therefore, more DFT coefficients used increases the accuracy and cost time for Euclidean distance computation.

(d) Yes, I did. The equation for calculating the Euclidean distance in complex number space is

$$\begin{aligned} d(\vec{u},\vec{v}) &= \|\vec{u}-\vec{v}\| \\ &= \sqrt{|u_1-v_1|^2+|u_2-v_2|^2+\cdots+|u_n-v_n|^2} \end{aligned}$$

Because $|(a+bi)-(c+di)|$ is equal to $|(a-bi)-(c-di)|$, therefore for calculating Euclidean distance I only checked from 1 to 1+ round((length of sequence-1)/2) and any DFT sequence which has a symmetric one I times the |ui - vi| by 2. Then I can save a lot loops (except first one and the mid one for odd length) when calculating the full Euclidean distance. And for part c, although I only used first 5,10 ,20 coefficients. But by using symmetric properties, I actually used 9,19, and 39 coefficients which should cost similar computation time but might increase the accuracy.
To compare the results, I also experiment the part(b) and part(c) without using symmetric property, the output as shown below.

ECG Dataset:

Num. of coefficients	**Accuracy (%)**	**Mean Cost time (ms)**	**Std for Cost time (ms)**
Early abandoning	92.49	0.007337	0.009189
5	91.71	0.003790	0.005345
10	92.31	0.004447	0.003040
20	92.31	0.005818	0.003280

NonInvasiveFatalECG_Thorax1 Dataset:

Num. of coefficients	**Accuracy (%)**	**Mean Cost time (ms)**	**Std for Cost time (ms)**
Early abandoning	82.90	0.007788	0.019178
5	73.99	0.004531	0.002586
10	81.32	0.005262	0.002891
20	82.90	0.006624	0.003271

Compare the table from part(b), while using Early abandoning strategy, the mean cost time has been reduced by using the symmetric property because of saving on the loop steps but keep the same accuracy level. Moreover, using the certain number of coefficients case, the mean cost time slightly been increased while using symmetric property that caused by the “times 2” computation, but there is no significant improvement in the accuracy.