

IE 582 Statistical Learning for Data Mining

Homework 5, due January 8th, 2020

Instructions: Please solve the following exercises using R (<http://www.r-project.org/>) or Python (<https://www.python.org/>). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
 - <https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/>
 - <https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41>

Things are little easier if you install Anaconda (<https://www.anaconda.com/>). Please export your work to an html file. Please provide your *.ipynb file in your repository and a link to this file in your html report will help us a lot.

- A Markdown html document. This can be created using RMarkdown for R and Python-Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself. As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Task

The conference paper by Liu et al. (2009) starts with the following statement: “Gestures have recently become attractive for spontaneous interaction with consumer electronics and mobile devices in the context of pervasive computing”. The aim is to provide efficient personalized gesture recognition on wide range of devices.

To achieve this, Liu et al. (2009) uses a single three-axis accelerometer to collect data from eight users to characterize eight gesture patterns. The library, uWaveGestureLibrary, consists over 4000 instances each of which has the accelerometer readings in three dimensions (i.e. x , y and z). Eight gestures are illustrated in Figure 1.

You are given the dataset on Moodle. There are separate files and each row corresponds to one gesture in the files. First column has the class information. The information between second and last column is the time ordered observations in the corresponding axis (provided in the file name as X, Y or Z).









1	2	3	4
			
5	6	7	8
			

Figure 1: Gesture vocabulary considered by Liu et al. (2009). The dot denotes the start and the arrow the end

- Read the data and visualize at least one instance (all axes) from each class and try to relate the shape (time series) you see with the gestures shown in Figure 1 (this is just for fun, sometimes it is good to start with data visualization to understand what is going on). A 3D scatter plot would be interesting. Note that this is not a position information. The acceleration information for each axis is provided in the files.
- Propose two distance measures for computing similarity between two time series. For each distance measure alternative, you are expected to evaluate the performance of two clustering approaches: k -means and hierarchical clustering. Since you are asked to work with distances, you need to use a modified version of k -means which is called the k -medoids approach. What is the difference between the k -means and k -medoids? In which condition(s) are they equivalent?
- For each distance measure alternative and clustering approach, you need to decide the number of clusters. Decide on the number of clusters to use for each clustering approach. Recall that there are alternative approaches to compute the distance between the clusters for hierarchical clustering. Consider at least two alternatives.
- Suppose, we decide to use the clustering results to do prediction on the test data. Report the accuracies for all the approaches you proposed. What is your conclusion?
- For the clustering approach that provides the best accuracy on the test data, draw the time series for each cluster on a scatter plot. In other words, you will have a time series plot for each cluster where the plot has all the time series from one cluster plotted over each other. Also, plot the average of the observations at each time point for each cluster on the same plot with different color. This can be interpreted as the cluster centroid which is expected to show the cluster prototype.

Reference

J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan. uWave: Accelerometer-based personalized gesture recognition and its applications. Pervasive Computing and Communications, IEEE International Conference on, 0:1-9, 2009.
(link: <http://www.ruf.rice.edu/~mobile/publications/liu09percom.pdf>)