DAT405 Introduction to Data Science and AI

2019-2010, Reading Period 1

Assignment 3: Clustering

There will be an overall grade for this assignment.

Some of the questions concern the main chain conformation of proteins. Part of a protein's main chain is shown in Figure 1. A protein chain is able to fold into its native conformation by rotation around two of the bonds in the main chain, designated ϕ (phi) and ψ (psi). Some combinations of phi and psi are impossible (e.g. some atoms clash into each other if we try to force the main chain to have a particular combination of phi and psi). Some other combinations of phi and psi are very common since they are energetically favourable.

The data files contain lists of phi and psi combinations that have been observed in proteins. The angles are measured here in degrees. The main file is "data_all.csv". Two smaller data files are provided ("data_200.csv" and "data_500.csv") and it might be convenient to test your programs quickly with these smaller files.

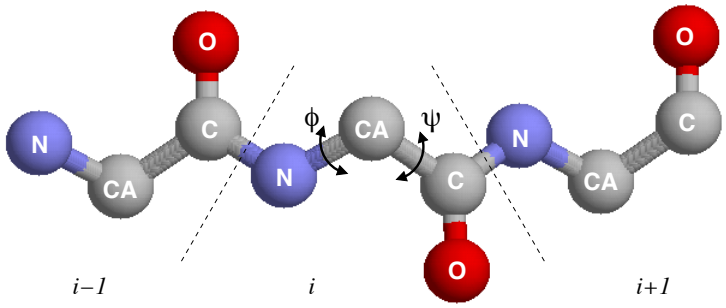

Figure 1. A protein's main chain. The heavy (i.e. non-hydrogen) main chain atoms of three consecutive amino acid residues (i-1, I and i+1) are represented by spheres, and the covalent bonds between these atoms are represented by rods. Nitrogen and oxygen atoms (N and O) are shown in blue and red respectively; carbon atoms are shown in grey. The central carbon atom (the alpha carbon, or Cα, labelled CA) is the main chain atom to which a side chain (not shown) is attached. Rotation can occur around the bonds labelled ϕ (phi) and ψ (psi).

1. Draw a scatter plot that shows the phi and psi combinations in the data file.

2. Use the K-means clustering method to cluster the phi and psi angle combinations in the data file.
    a. Select a suitable distance metric for this task. If this is different from the Euclidean distance function, explain how it differs. For a higher grade, motivate the choice of distance metric.
    b. Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.
    c. Validate the clusters that are found with the chosen value of K.

3. Use the DBSCAN method to cluster the phi and psi angle combinations in the data file.
    a. Motivate:
        i. the choice of the minimum number of samples in the neighbourhood for a point to be considered as a core point, and
        ii. the choice of the maximum distance between two samples belonging to the same neighbourhood (“eps” or “epsilon”).
    b. Highlight the clusters found using DBSCAN and any outliers in a scatter plot. How many outliers are found? Plot a histogram to show which amino acid residue types are most frequently outliers.
    c. Compare the clusters found by DBSCAN with those found using K-means.
    d. Discuss whether the clusters found using DBSCAN are robust to small changes in the minimum number of samples in the neighbourhood for a point to be considered as a core point, and/or the choice of the maximum distance between two samples belonging to the same neighbourhood (“eps” or “epsilon”).

4. The data file can be stratified by amino acid residue type. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters. Similarly, investigate how the clusters found for amino acid residues of type GLY differ from the general clusters. Remember that parameters might have to be adjusted from those used in previous questions.

Submitting work

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment.

If you upload a zip file, please also upload any PDF files separately (so that they can be viewed more conveniently in Canvas).

Deadline: Monday 23 September 2019 at 12:00 (noon).