CSGE603130 • Kecerdasan Artifisial dan Sains Data Dasar
Semester Ganjil 2021/2022
Fakultas Ilmu Komputer, Universitas Indonesia

Tugas 2: *Clustering*

Tenggat Waktu: 12 Oktober 2021, 23.55 WIB

Ketentuan:

1. Dataset yang digunakan pada tugas ini beserta deskripsinya telah disediakan di SCeLe.
2. Buatlah program Jupyter Notebook yang menjawab pertanyaan sesuai dengan perintah soal yang disediakan.
3. Program Jupyter Notebook yang telah dibuat dikumpulkan dengan format penamaan **Kelas_TugasX_NPM_Nama.ipynb**
   Contoh: F_Tugas2_1706979341_Lulu Ilmaknun Qurotaini.ipynb
4. Kumpulkan dokumen tersebut pada submisi yang telah disediakan di SCeLe sesuai dengan kelas masing-masing sebelum **12 Oktober 2021, 23.55 WIB**. Keterlambatan pengumpulan akan dikenakan pinalti.
5. Tugas ini dirancang sebagai **tugas mandiri**. Plagiarisme tidak diperkenankan dalam bentuk apapun. Adapun kolaborasi berupa diskusi (tanpa menyalin maupun mengambil jawaban orang lain) dan literasi masih diperbolehkan dengan mencantumkan kolaborator dan sumber.
6. Template untuk pengerjaan Tugas 2 dapat diakses pada link berikut:
   https://drive.google.com/file/d/1ssciBR7-zC3l2177hz5E7nCitRVisE3F/view?usp=sharing
   credit to Muhammad Aulia Adil Murtito (2019)

Soal Tugas 2

Catatan: Algoritma clustering yang boleh dipakai pada tugas ini hanya K-means dan Agglomerative.

Soal 1 [15 Poin] - Teori

a. Jelaskan mengapa K-means clustering tidak cocok untuk yang bukan *hyper-spheres*?
b. Jelaskan mengapa algoritma Hierarchical clustering dapat digunakan tanpa perlu menetapkan jumlah *cluster*?
c. Jelaskan apa itu nilai metrik intra-class *similarity* dan cara menghitungnya!
d. Jelaskan apa itu nilai metrik inter-class *similarity* dan cara menghitungnya!

Soal 2 [40 Poin] - Guess the clustering

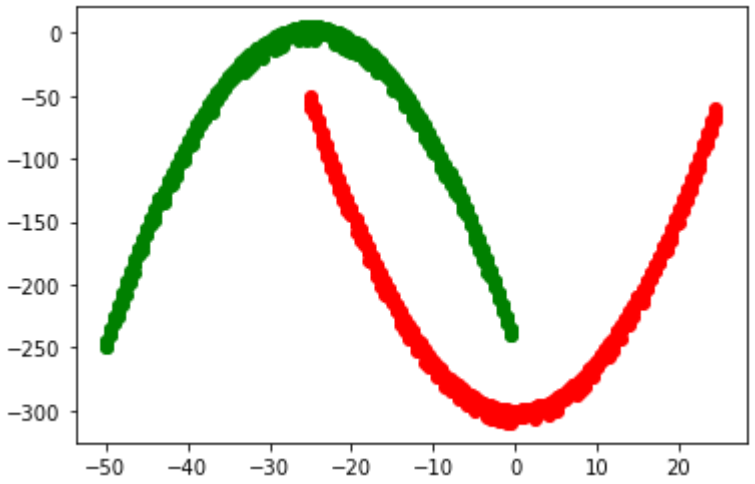
a. Buka data dari soal2.csv dan plot semua titik di plot dua dimensi!
b. Dari plot tersebut, usulkan jumlah klaster yang dapat membagi data dengan baik!
c. Gunakan sebuah algoritma *clustering* yang dapat membagi data sesuai dengan jumlah klaster yang Anda usulkan. Jelaskan mengapa algoritma *clustering* tersebut digunakan!

   Catatan: Jawaban seperti karena "K-Means bisa/baik dipakai untuk clustering" tidak cukup. Semua algoritma *clustering* bisa dipakai untuk *clustering*. Tapi kenapa itu?
d. Visualisasikan hasil *clustering* dengan menampilkan scatter plot data yang di *color-coded* berdasarkan klasternya. Selain itu, cetak jumlah *cluster* yang dihasilkan!

   Hint: Pelajari parameter *c* pada fungsi matplotlib.pyplot.scatter()
e. Hitung nilai *intra-class similarity* hasil clustering tersebut dengan menghitung jumlah dari jarak (*sum of distance*) masing-masing sampel dengan pusat klasternya! Koordinat sebuah pusat klaster adalah rata-rata dari semua sampel di klaster tersebut. Rumus jarak yang digunakan adalah L2-norm/*Euclidean Distance*.
f. Hitung nilai *silhouette coefficient* dari hasil *clustering* tersebut!

Soal 3 [30 Poin] - Hierarchical Clustering

a. Buka data dari soal3.csv dan plot semua titik di plot dua dimensi
b. Pada soal ini, anda diharapkan melakukan *clustering* menggunakan algoritma Agglomerative sehingga terbuat klaster seperti berikut:



c. Kemungkinan besar saat Anda menggunakan modul algoritma Agglomerative dari library sklearn, Anda harus mengubah parameter yang digunakan. Sebutkan parameter apa saja yang

anda gunakan beserta nilainya, dan jelaskan alasan anda menggunakan parameter tersebut dan menggunakan nilai tersebut.

d. (Bonus) Lakukan *clustering* menggunakan K-Means pada data tersebut, visualisasikan hasilnya, kemudian lakukan analisis terhadap hasil tersebut.

Soal 4 [20 Poin] - Which came first: dimensionality reduction or clustering?

a. Buka data dari soal4.csv dan tampilkan 10 baris pertama! Hitung jumlah data dan jumlah fitur!
b. Ikuti langkah-langkah berikut:
    i. Salin data asli dan masukkan dalam variable *data_copy_1*!
    ii. Lakukan *Dimensionality reduction (PCA)* pada *data_copy_1* sehingga jumlah fiturnya menjadi 2!
    iii. Visualisasikan *data_copy_1* yang sudah direduksi, kemudian tentukan berapa jumlah cluster yang tepat!
    iv. Lakukan clustering pada *data_copy_1* yang sudah direduksi menggunakan K-Means dengan parameter `random_state`=2021!
    v. Visualisasikan hasil clustering pada *data_copy_1* yang sudah direduksi!
c. Ikuti langkah-langkah berikut:
    i. Salin data asli dan masukkan dalam variable *data_copy_2*!
    ii. Lakukan clustering pada *data_copy_2* menggunakan K-Means dengan parameter `random_state`=2021 dan jumlah cluster sama dengan jumlah cluster yang kamu gunakan pada poin b!
    iii. Lakukan *Dimensionality reduction (PCA)* pada *data_copy_2* sehingga jumlah fiturnya menjadi 2!
    iv. Visualisasikan hasil clustering pada *data_copy_2* yang sudah direduksi!
d. Apakah ada perbedaan pada hasil poin b dan poin c? Mengapa demikian? Apa kesimpulan yang bisa anda ambil?