Università degli Studi di Palermo

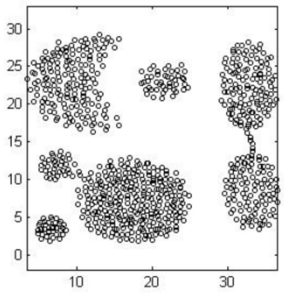
Dipartimento di Ingegneria
Direttore: prof. Antonino Valenza

dipartimento di ingegneria unipa

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

Prova scritta del 29 giugno 2023

Si consideri il data set `Aggregation.txt`, riportato in figura e allegato al presente compito, costituito da 788 punti sintetici bidimensionali appartenenti a 7 cluster e affetti da single linkage:



1. Ricercare gli outlier utilizzando il metodo *Local Outlier Factor* (LOF) e stabilire autonomamente il numero di vicini da considerare nell'intorno di ogni punto, testando almeno tre valori di *k*.

   punti ___/ 8

2. Eseguire il clustering con l'algoritmo DBSCAN testando almeno tre diverse combinazioni di coppie {*Eps*, *MinPoints*}.

   punti ___/ 8

3. Implementare una routine routine di clustering probabilistico con approccio EM, usando una mistura di 7 distribuzioni gaussiane bivariate in cui i valori iniziali dei prior siano tutti pari a $\alpha_i = \frac{1}{7}$.

   punti ___/ 10

4. Determinare e mostrare la matrice di confusione di tutti gli algoritmi di clustering implementati (tutti i tentativi DBSCAN e il clustering probabilistico) calcolando i relativi indici di Gini.

   punti ___/ 4

TOTALE: punti ___/ 30

Regole della prova scritta

Di seguito si riportano le regole da seguire e le caratteristiche della prova ai fini della valutazione:

1. La durata complessiva della prova è pari a due ore e prevede una serie di quesiti che approfondiscono diversi aspetti dello stesso problema: ognuno sarà libero di dedicare ad ogni quesito il tempo che vorrà.
2. La prova si svolge ***interamente*** al calcolatore.
3. L'ambiente di sviluppo predisposto è un environmet conda/mamba dotato di editor Spyder e dei seguenti pacchetti:
   a. Pandas
   b. Matplotlib
   c. Seaborn
   d. Numpy
   e. Scipy
   f. Scikit-learn
   g. Pyspark
   h. Keras
   i. Tensorflow

Data: _____ Allievo: _____ Matricola: _____

4. Ai fini dell'avvio dell'ambiente, aprire il prompt dei comandi e digitare i due seguenti comandi:
```
$ mamba activate spyder-env
$ spyder
```

5. Sarà consentito consegnare dopo la prima ora di prova.
6. Sarà necessario spegnere e consegnare i dispositivi mobili (smartphone, smartwatch e tablet) alla cattedra prima dello svolgimento della prova.
7. La navigazione internet dalle postazioni sarà bloccata, in generale, e consentita solo verso i siti di documentazione delle librerie ed il repository delle slide in pdf.
8. Il docente distribuirà copia digitale del compito ed eventuali data set direttamente dalla propria postazione ovvero tramite penna USB e allo stesso modo raccoglierà gli elaborati di programmazione.
9. Il candidato consegnerà comunque il presente foglio datato e con l'indicazione del nome e del numero di matricola.
10. Ai fini del calcolo del voto finale della prova, il *valore massimo* di ciascun quesito è riportato in calce allo stesso. La prova riceverà una valutazione pari alla somma dei voti riportati in ciascun quesito. Si precisa che il docente attribuirà ad ogni quesito una votazione *non binaria*, cioè non tutto il valore oppure 0, ma valuterà la correttezza formale dell'elaborato, il rigore metodologico dell'approccio teorico e l'originalità delle soluzioni proposte per attribuire una votazione nel range definito dal valor massimo del quesito.

Data: _____ Allievo: _____ Matricola: _____