# UCLA CS 145 Homework #2

DUE DATE: Monday May 3, 23:59 PM

## Note

- Late submissions will generally **NOT** be accepted. Each student has an one-day extension for **ONE** of the three homework assignments if the student contacts the instructor and TA **BEFORE** the submission deadline to arrange the only late submission.
- Discussions on homework assignments are encouraged, but any form of cheating and plagiarism will **NOT** be tolerated. Every student must submit his/her own solutions on Gradescope by the deadline. *Suspicious cases will be reported to The Office of the Dean of Students.*

## 1 K-Means

(a) (10%) Given the following points in a 2-dimensional space:

$$(1,1),(1,2),(2,0),(2,2),(2,3),(2,4),(3,3),(4,2),(4,4)$$

Simulate K-means with K=2 for 3 iterations on the paper and show the result after each iteration (a table of clustering result is fine here). Pick $(1,1)$ and $(4,4)$ as your initial centroids.

(b) (10%) Prove the convergence of the K-means algorithm. In other words, prove that K-means algorithm converges in finite iterations.

## 2 Comparison of Clustering Algorithms

(a) (10%) List the pros and cons for (1) K-Means (2) DBSCAN and (3) BIRCH (At least one pro and two cons for each algorithm to get full marks)

(b) (10%) Consider applying (1) K-Means and (2) DBSCAN on two different datasets as shown in Figure 1. Assume that the ground-truth numbers of clusters in two datasets are 3 and 2 respectively, indicated by the blue circles in Figure 1(a) and (c). The clustering results for these two algorithms are shown in Figure 1 using different colors. Explain what Algorithm 1 and Algorithm 2 could be with detailed reasons.

(c) (10%) Currently, Algorithm 2 has divided the upper cluster of Data 2 into two different clusters as denoted by the green and yellow groups in Figure 1(d). Is there any possible way to avoid it such that the upper cluster can be clustered as a whole by Algorithm 2? If so, what would be your solution?

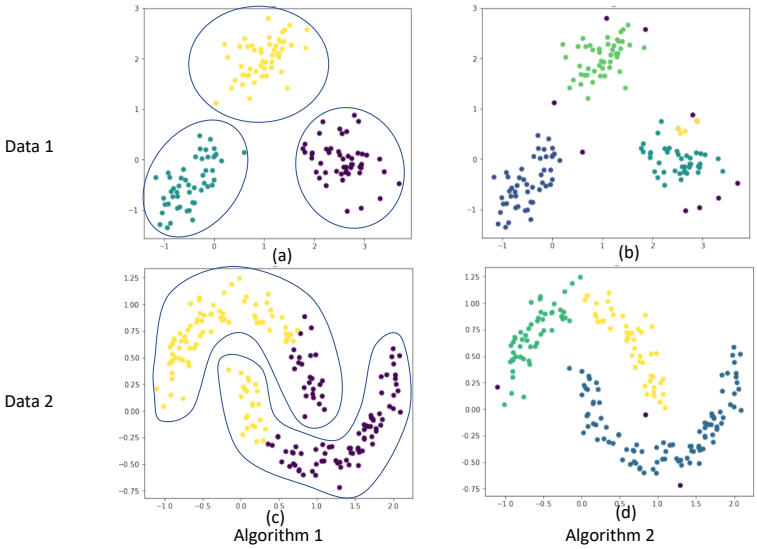


Figure 1: Clustering results for DBSCAN and KMeans respectively. Color denotes clusters generated by algorithms, blue circle denotes the ground truth clusters (data 1 has 3 clusters and data 2 has 2 clusters)

## 3 Decision Tree

Suppose that our dataset includes a total of 800 people with 400 males and 400 females, and our goal is to do gender classification. Consider two different possible attributes we can split on in a decision tree model. Split on the first attribute results in a node11 with 300 males and 100 females, and a node12 with 100 males and 300 females. Split on the second attribute results in a node21 with 400 males and 200 females, and a node22 with 200 females only.

(a) (5%) What is the entropy in each of these four nodes?

(b) (10%) What is the information gain of each of the two splits? Which split do you prefer if the measurement is information gain.

(c) (10%) What is the gain ratio (normalized information gain) of each of the two splits? Which split do you prefer under this measurement. Do you get the same conclusion as information gain?

## 4 Naïve Bayes Classifier

(25%) For the data shown in Table 1, build a naïve Bayes classifier and predict the probability of each training sample belonging to each class. Then use the classifier to classify a testing sample $(F_1 = Y, F_2 = N, F_3 = N)$. Note that the details of the computational process are needed. More specifically, you need to show all of the conditional probabilities within computations.

| Class | Feature 1 ($F_1$) | Feature 2 ($F_2$) | Feature 3 ($F_3$) |
|---|---|---|---|
| Class A | N | N | Y |
| Class A | N | Y | Y |
| Class A | N | Y | Y |
| Class A | N | Y | Y |
| Class A | N | Y | Y |
| Class A | Y | N | N |
| Class A | Y | N | Y |
| Class A | Y | Y | Y |
| Class A | Y | Y | Y |
| Class B | N | N | N |
| Class B | N | N | N |
| Class B | N | N | N |
| Class B | N | Y | Y |
| Class B | Y | N | N |
| Class B | Y | N | Y |
| Class B | Y | Y | N |
| Class B | Y | N | N |
| Class B | Y | N | N |
| Class B | Y | N | N |

Table 1: Training data for Question 4.