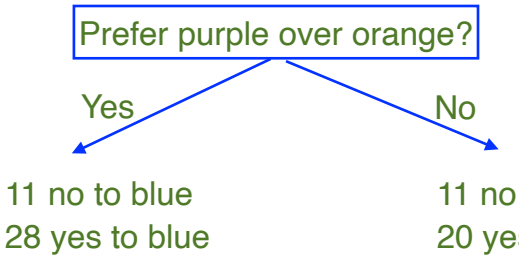
CSE514, Fall 2022, HW 5 Name: Student ID:
Note: This homework is worth a total of 15 points

Please refer to the
Color preferences.csv
file posted on Canvas to find the probability values needed for this first page of questions.

Q1 (9pts): Create a stump decision tree by splitting the data based on purple preference, where the left child holds all the samples that answered Yes, and the right child holds all the samples that answered No.

Q1a: Given the task of predicting blue preference, what's the Total Weighted Gini Impurity of this split?



GI_yes = 1-[(11/39)^2 + (28/39)^2]=0.405

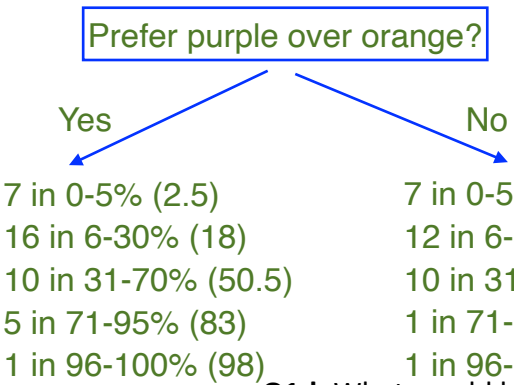
GI_no = 1-[(11/31)^2 + (20/31)^2]=0.458

TWGI = 39/70*GI_yes + 31/70*GI_no=0.428

Q1b: What would be the predicted labels of the left and right leaves?

Both leaves predict yes in response to preferring blue over red

Q1c: Given the task of predicting percentage of white t-shirts, what's the Total Weighted Variance of this split?



Var_yes = Mean[(sample-Mean_yes)^2]=Mean[(sample-33.9)^2]
= 765.8 or 0.0766

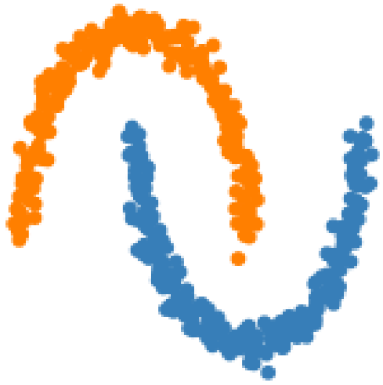
Var_no = Mean[(sample-Mean_no)^2]=Mean[(sample-29.7)^2]
= 601.7 or 0.0602

TWVar = 39/70*Var_yes + 31/70*Var_no=693.1 or 0.0693

Q1d: What would be the predicted values of the left and right leaves?

Left predicts 33.9%, Right predicts 29.7%

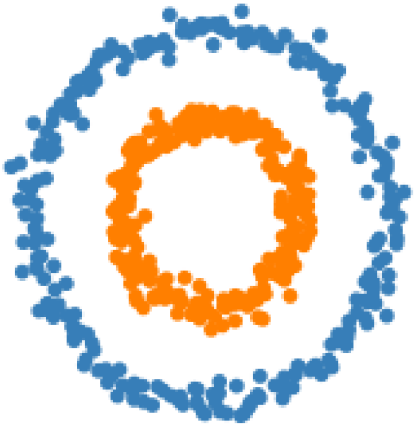
Q2 (6pts): For each of the following clusters, give a clustering method you would NOT recommend, and a reason why you think it would fail.
Don't repeat clustering methods



Q2a:

k-means
centroid/complete-linkage hierarchical/agglomerative
Mean Shift
EM with GMM

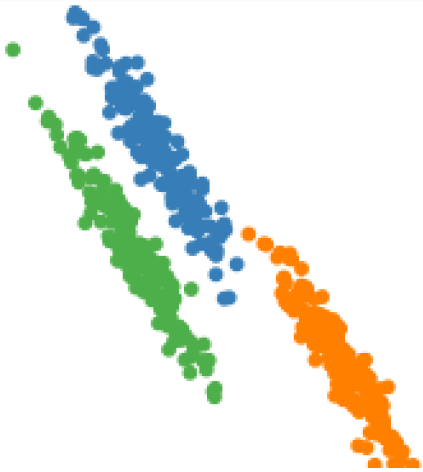
Will all fail because the data is not spherical and/or not linearly separable



Q2b:

k-means
centroid/complete-linkage hierarchical/agglomerative
Mean Shift
EM with GMM

Will all fail because the data is not linearly separable



Q2c:

k-means
centroid/complete-linkage hierarchical/agglomerative
MeanShift

Will all fail because the clusters are not spherical

DBScan could fail because the low density areas should not be outliers