ECE M148 Homework 5
Introduction to Data Science Due: May 26, 12:00 PM
Instructor: Lara Dolecek TAs: Harish GV, Jayanth Shreekumar

Please upload your homework to Gradescope by May 26, 12:00 PM. You can access Gradescope directly or using the link provided on BruinLearn. You may type your homework or scan your handwritten version. Make sure all the work is discernible.

1. Consider the following dataset which shows the different characteristics of each day and whether I played tennis or not:

Day	Humidity	Wind	Play Tennis
1	High	Weak	Yes
2	High	Strong	No
3	Normal	Weak	Yes
4	Normal	Weak	Yes
5	High	Strong	No
6	Normal	Strong	Yes
7	High	Weak	Yes
8	Normal	Weak	Yes
9	High	Strong	Yes
10	High	Strong	No
11	High	Weak	Yes
12	High	Weak	No
13	High	Strong	No
14	Normal	Strong	No

Suppose we wish to use a decision tree to predict whether I play tennis or not.

(a) Calculate the Gini Index and Gini Index Gain for each feature split (Humidity or Wind).

(b) What feature provides the best Gini Index Gain?

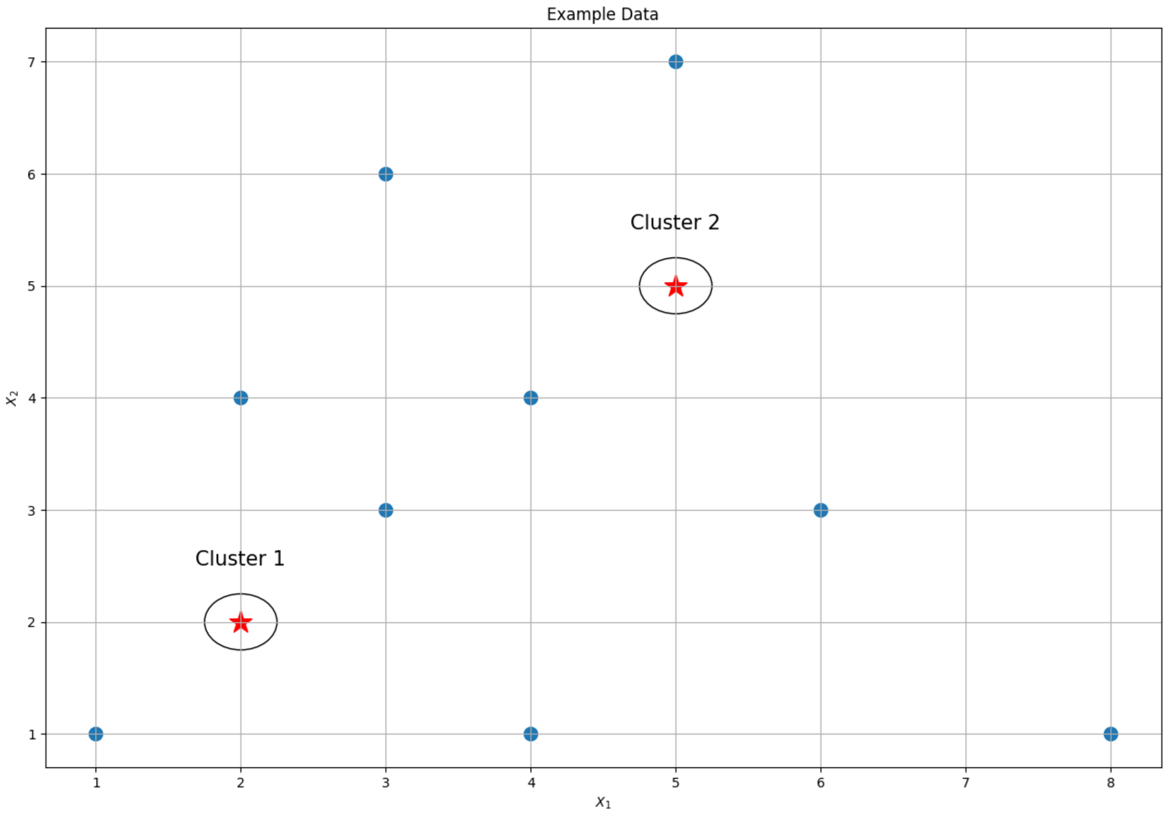
(c) Now, use the entropy function discussed in class. Afterward, calculate the Information Gain using entropy. Note that all entropy calculations should use logarithms in base 2.

(d) Does using entropy over Gini change the best feature? If so, what is the new best feature split?

2. Consider the following dataset:

Sample	X_1	X_2
1	1	1
2	2	4
3	4	1
4	6	3
5	5	7
6	8	1
7	4	4
8	3	6
9	3	3

We will use K-means to cluster this data. Assume that we initialize cluster 1 centers at $[2, 2]$ and cluster 2 center at $[5, 5]$. We can see the centers and data on the following plot:

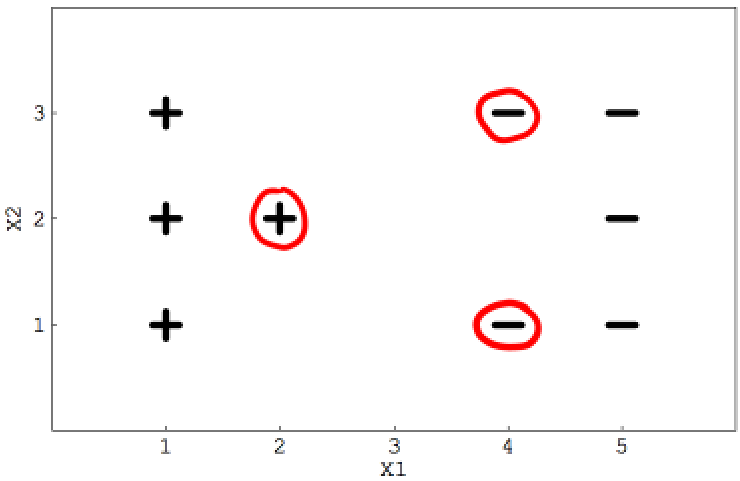


Perform one iteration of K-means clustering by

- Assigning each data sample to the cluster with the closest mean.
- Getting the new cluster center by averaging all the points within the cluster.

Show all your work. Your final answer should be the new cluster centers and which cluster each sample data point belongs to.

3.



(a) Consider the pictured dataset with 2 classes ('+' and '-'). If you remove one of the points that **is not** circled, how will this affect the decision boundary of an SVM?

(b) What is the difference between a hard margin and soft margin SVM?

(c) If we remove the sample related to the circled "+" and run a hard margin SVM, how many support vectors will the algorithm determine? Justify your answer.

4. Show that the following representations of the probabilities of a test point X belonging to a class $Y = i$ are equivalent in logistic regression:

(a)

$$P(Y = i|X) = \frac{e^{\beta_{0i}+\beta_{1i}X}}{1 + \sum_{j=1}^{K-1} e^{\beta_{0j}+\beta_{1j}X}}, 1 \leq i \leq K - 1$$

$$P(Y = i|X) = \frac{e^{\tilde{\beta_{0i}}+\tilde{\beta_{1i}}X}}{\sum_{j=1}^{K} e^{\tilde{\beta_{0j}}+\tilde{\beta_{1j}}X}}, 1 \leq i \leq K$$

(b) Given $X = 5, K = 3$ and the following $\tilde{\beta}$ values found during training:

Class i	$\tilde{\beta_{0i}}$	$\tilde{\beta_{1i}}$
1	-0.2	0.06
2	0.2	0.04
3	0.3	0.5

Which class does the test point X get assigned?

5. True or False questions. For each statement, decide whether the statement is True or False and provide justification (full credit for the correct justification).

   (a) For regression trees, we pick the feature whose split maximizes the MSE.

   (b) K-means will always converge to the same solution regardless of initial points chosen for the means.

   (c) Agglomerative clustering is the process of combining clusters together in order to minimize the overall distortion.

   (d) In soft margin SVM, larger constant λ for the slack variables implies wider margin for the training data.

   (e) The purpose of using a random forest of shallow decision trees learned on bootstrapped samples versus a single deep decision tree learned on the whole dataset is to avoid overfitting.