ECE M148
Introduction to Data Science
Instructor: Lara Dolecek

Homework 5 Solutions
Due: May 26, 12:00 PM
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?

Solution:

(a) Let $G(Before)$ be the gini index before any split occurs.

$$G(Before) = 1 - (\frac{6}{14})^2 - (\frac{8}{14})^2 \approx 0.4897$$

.

Gini index for each split on *Humidity*:

$$G(High) = 1 - (\frac{4}{9})^2 - (\frac{5}{9})^2 = 0.493827$$
$$G(Normal) = 1 - (\frac{4}{5})^2 - (\frac{1}{5})^2 = 0.32$$

Average Gini index for *Humidity*:

$$G(Humidity) = \frac{9}{14}G(High) + \frac{5}{14}G(Normal) \approx 0.4317459$$

Gini Index Gain for *Humidity*:

$$Gain(Humidity) = G(Before) - G(Humidity) \approx 0.0579541$$

Gini index for each split on *Wind*:

$$G(Weak) = 1 - (\frac{6}{7})^2 - (\frac{1}{7})^2 = 0.24489$$
$$G(Strong) = 1 - (\frac{5}{7})^2 - (\frac{2}{7})^2 = 0.408162$$

Average Gini index for *Wind*:

$$G(Wind) = \frac{7}{14}G(Weak) + \frac{7}{14}G(Strong) \approx 0.326526$$

Gini Index Gain for *Wind*:

$$Gain(Wind) = G(Before) - G(Wind) = 0.4897 - 0.326526 \approx 0.163174$$

(b) From the calculations in part (a), we see that the best feature to split on is Wind.

(c) Recall that entropy is calculated as $H([p_i]_{i=1}^n) = -\sum_{i=1}^n p_i \log_2(p_i)$. We use the term E to denote the entropy of splitting on a particular feature value and the average entropy of splitting on a feature.

The entropy before any split occurs is $H([\frac{6}{14}, \frac{8}{14}]) = 0.985228$.

Entropy for each split on *Humidity*:

$$E(High) = H([\frac{4}{9}, \frac{5}{9}]) = 0.991076$$
$$E(Normal) = H([\frac{4}{5}, \frac{1}{5}]) = 0.721928$$

Average entropy of splitting using *Humidity*::

$$E(Humidity) = \frac{9}{14}E(High) + \frac{5}{14}E(Normal) \approx 0.89495$$

Information Gain for *Humidity*::

$$Gain(Humidity) = 0.985228 - 0.89495 = 0.090276$$

Entropy for each split on *Wind*:

$$E(Weak) = H([\frac{6}{7}, \frac{1}{7}]) \approx 0.59167$$
$$E(Strong) = H([\frac{5}{7}, \frac{2}{7}]) = 0.86312$$

Average entropy of splitting using *Wind*:

$$E(Wind) = \frac{7}{14}E(Weak) + \frac{7}{14}E(Strong) \approx 0.727395$$

Information Gain for *Wind*:

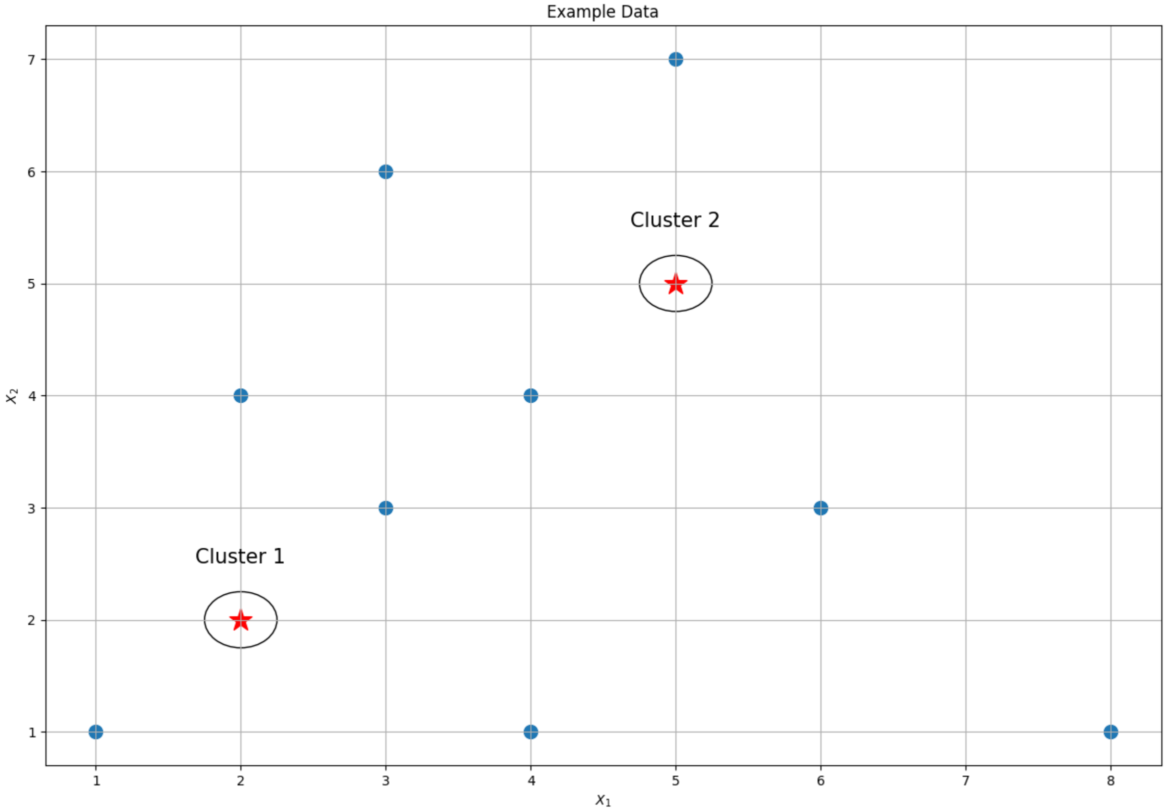
$$Gain(Wind) = 0.985228 - 0.727395 \approx 0.257833$$

(d) Using entropy did not change the best feature for splitting which is Wind.

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.

Solution:

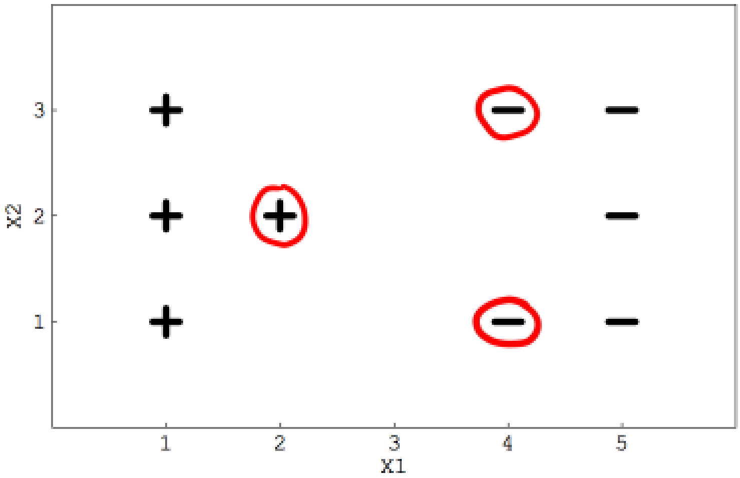
First step: We assign samples 1, 2, 3, and 9 to cluster 1 since this cluster is closest to these points. Similarly, we assign samples 4, 5, 6, 7, and 8 to cluster 2.

Second step:

New mean of cluster 1 : $[X_1, X_2] = \begin{bmatrix} 2.5 & 2.25 \end{bmatrix}$

New mean of cluster 2 : $[X_1, X_2] = \begin{bmatrix} 5.2 & 4.2 \end{bmatrix}$

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.

Solution:

(a) It will not affect the decision boundary because the circled points are the closest to the decision boundary which is located at $X1 = 3$.

(b) Hard Margin SVM tries to find a separating hyperplane for the datasets and only works if the data is linearly separable. Soft Margin SVM allows some misclassification of points and aims to minimize the distance these points have from the margin.

(c) There would be 5 support vectors since the decision boundary can now move leftward until the 3 points on the left are on the margin. Thus, the 3 points on the left and the circled "-" would be the support vectors.

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?

Solution

(a) Starting with the following probability from the second set (softmax equations), we have:

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

Pulling out $e^{\tilde{\beta_{0K}}+\tilde{\beta_{1K}}X}$ in the denominator and separating the K^{th} term from the summation,

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

Cancelling similar terms and renaming $\tilde{\beta_{0j}} - \tilde{\beta_{0K}}$ as β_{0j} and $\tilde{\beta_{1j}} - \tilde{\beta_{1K}}$ as β_{1j}:

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

Similarly for the other terms, we have:

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

Taking the exponent term to the numerator and renaming just as before, we get:

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

for all $k \neq K$.

Hence proved.

(b) Using the softmax formula, we compute the probabilities for each class:

$$\begin{aligned}
\text{Denominator} &= e^{-0.2+0.06\times 5} + e^{0.2+0.04\times 5} + e^{0.3+0.5\times 5} = 19.041642 \\
\text{Numerator Class 1} &= e^{-0.2+0.06\times 5} = 1.10517091808 \\
\text{Numerator Class 2} &= e^{0.2+0.04\times 5} = 1.49182469764 \\
\text{Numerator Class 3} &= e^{0.3+0.5\times 5} = 16.4446467711
\end{aligned}$$

Then, the probabilities are:

$$\begin{aligned}
\text{Class 1} &= \frac{e^{-0.2+0.06\times 5}}{\text{Denominator}} = 0.05803968 \\
\text{Class 2} &= \frac{e^{0.2+0.04\times 5}}{\text{Denominator}} = 0.078345 \\
\text{Class 3} &= \frac{e^{0.3+0.5\times 5}}{\text{Denominator}} = 0.8636149
\end{aligned}$$

Therefore, the test point X is classified as class 3.

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.

   Solution:

   (a) False. Each split is chosen to minimize the MSE.

   (b) False. K-means is highly sensitive to the choice of initial means.

   (c) True. Agglomerative clustering starts out with small clusters and combines them in order to reduce the distortion.

   (d) False. Since λ is the term associated with the margin error, increasing λ tells the model to prioritize the mis-classification penalty over the margin. Thus, the margin should generally get smaller for the training data.

   (e) True. Since each shallow decision tree learns on a bootstrapped sample, the effect of variance is averaged out by the random forest. Thus, random forests are less likely to overfit.