CIS 520, Machine Learning, Fall 2021
Homework 1
Due: Sunday, September 19th, 11:59pm
Submit to Gradescope

Yixuan Meng, Zhouyang Fang

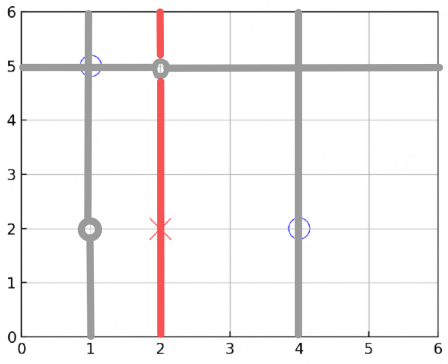
1 Non-Normal Norms

1. For the given vectors, the point closest to x_1 under each of the following norms is

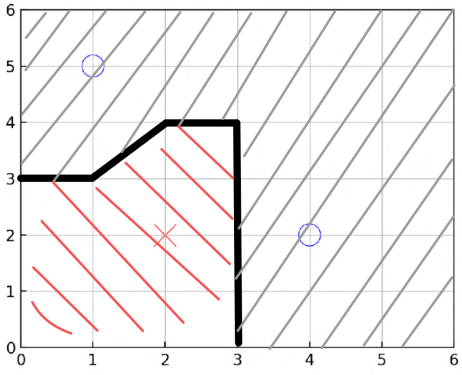
   a) L_0: x_3 with distance = 3.0
   b) L_1: x_3 with distance = 9.5
   c) L_2: x_4 with distance = 5.3
   d) L_{inf}: x_4 with distance = 4.2

2. Draw the 1-Nearest Neighbor decision boundaries with the given norms and lightly shade the o region:

   a) L_0

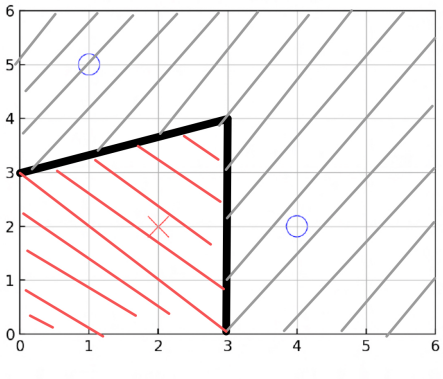
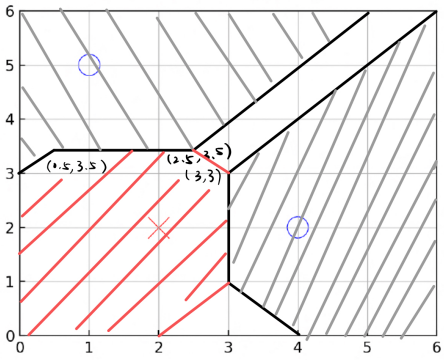


   b) L_1



   c) L_2

   d) L_{inf}

2 Decision trees

1. Information gain and KL-divergence.

   (a) If variables X and Y are independent, is $IG(x, y) = 0$? If yes, prove it. If no, give a counter example.

   $$IG(x, y) = 0$$

   Since $P(X, Y) = P(X)P(Y|X)$, and X and Y are independent, then $P(X, Y) = P(X)P(Y)$, therefore in KL-divergence's formula, $log(\frac{p(x)p(y)}{p(x,y)}) = 0$

   (b) Prove that $IG(x, y) = H[x] - H[x \mid y] = H[y] - H[y \mid x]$, starting from the definition in terms of KL-divergence:

$$\begin{aligned}
IG(x, y) &= KL\left(p(x, y)||p(x)p(y)\right) \\
&= \sum_x \sum_y p(x, y) log(\frac{p(x, y)}{p(x)p(y)}) \\
&= \sum_x \sum_y p(x, y) log(\frac{1}{p(x)}) + \sum_x \sum_y p(x, y) log(\frac{p(x, y)}{p(y)}) \\
&= -\sum_x p(x) log(p(x)) - (-\sum_x \sum_y p(x, y) log(\frac{p(x, y)}{p(y)})) \\
&= H(x) - (\sum_y p(y)(-\sum_x P(x|y) log(x|y))) \\
&= H[x] - H[x \mid y]
\end{aligned}$$

   Similarly, $IG(x, y) = H[y] - H[y \mid x]$

3 K-nearest neighbors Classification (Programming)

1. How does having a larger dataset might influence the performance of KNN?

   A larger dataset can dramatically increase the complexity of KNN, because the algorithm need to iterate all data points and calculate distances. A large dataset can also improve accuracy of KNN by overfitting on small number of data samples.

2. Tabulate your results in Table 1 for the **validation set**.

K	**Norm**	**Accuracy (%)**
3	L1	72.17
3	L2	69.57
3	L-inf	73.04
5	L1	75.65
5	L2	76.52
5	L-inf	72.17
7	L1	73.04
7	L2	77.39
7	L-inf	72.17

Table 1: Accuracy for the KNN classification problem on the validation set

3. Finally, mention the best K and the norm combination you have settled upon from the above table and report the accuracy on the test set using that combination.

   The best combination is K = 7 with L2 norm, the accuracy on test set using this combination is 71.43%.

4 Decision Trees (Programming)

4.1 Part 1: Effects of Dataset Size on Performance

1. Report the training, validation, and test accuracies on the full and partial datasets below. Note that this portion will be graded by the Autograder.

Accuracy Scores		
	Full Dataset	Small Dataset
Training Accuracy	1.0000	1.0000
Validation Accuracy	0.7043	0.7130
Test Accuracy	0.7532	0.6753

2. Which dataset had a higher difference between training and test accuracy? Briefly explain why.

   The Full Dataset. Because the tree overfit to the train data, and does not generalize well to fit test data.

4.2 Part 2: Effects of Dataset Size on Performance

1. Report the chosen hyperparameters for the complete and partial set below. Note that this section will be graded by the Autograder.

Grid Search Chosen Hyperparameters		
	Full Dataset	Small Dataset
Tree Depth	3	1
Max Leaf Nodes	4	2

2. Did the small dataset have higher or lower chosen hyperparameter values than the full dataset? Briefly explain why.

   The smaller dataset hace lower chosen hyperparameters. For smaller dataset, a decision tree with smaller depth (shorter) and leaf counts (less splits) would be less likely to overfitting,

4.3 Part 3: Retrain Decision Tree and Plot Hyperparameter Search

1. Report the train, validation, and test accuracies after retraining the decision tree with the new hyperparameters. Also paste in the values for the training and validation scores lists when varying the max leaf node count hyperparameter.

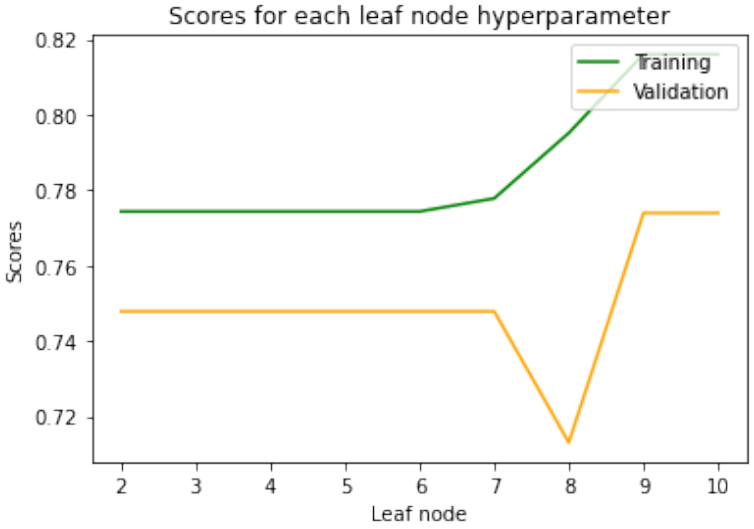
Retrained Decision Tree Performance for Small Dataset	
	Score
Training Accuracy	0.8160
Validation Accuracy	0.7739
Test Accuracy	0.7143

Training and Validation List Values	
	List
Training	[0.7743, 0.7743, 0.7743, 0.7743, 0.7743, 0.7778, 0.7951, 0.8160, 0.8160]
Validation	[0.7478, 0.7478, 0.7478, 0.7478, 0.7478, 0.7478, 0.7130, 0.7739, 0.7739]

2. How did the training accuracy and testing accuracy change after tuning compared to before? Briefly explain why.

   The training accuracy decreases, but the testing accuracy improves.
   Because during the grid search, the validation accuracy is used as objective to choose best hyperparameter, helping to mitigate overfitting on training set, so the model can generalize better.

3. Paste the plot of training and validation scores with different leaf count values on the small dataset. Explain any trends or patterns with the plot within validation and training scores and briefly explain why.



   The train accuracy score keeps increasing, while the validation score fluctuates. That is because as leaf node increase, more and more splits are created, the model starts to overfit, and produce random wrong labels. When some of the random classifications give better guesses, the validation score increases.

5 Feature Scaling Effects (Programming)

1. Report the training and testing accuracies for unstandardized and standardized data for both Decision Trees and KNNs using their default hyperparameter values.

Scores for Unstandardized and Standardized Data				
	KNN Unscaled	KNN Scaled	DT Unscaled	DT Scaled
Training Accuracy	0.7899	0.8177	1.0000	1.0000
Test Accuracy	0.7013	0.8182	0.7532	0.7532

2. What happens to performance when we use standardization for data with decision trees? What about KNN? Briefly explain why each happened.

   After standardizing input data, the accuracy of the decision tree didn't change, but the accuracy of KNN is improved.
   It is because the decision tree is based on entropy, and the data variance would not affect its performance, whereas KNN is based on distance, data normalization would make sure that each feature (input data cols) affect the model proportionally.