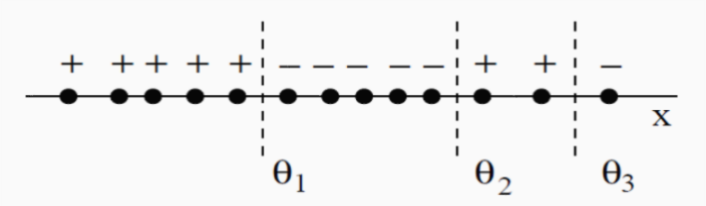
CSEN 240 Machine Learning
Homework #3

(Level 1: high-level understanding)

1. (14 points) Which of the following can describe Decision Tree? (select all that apply)
    a. It is a supervised learning algorithm.
    b. It can be used to solve a classification problem.
    c. It is a tree-like structure that consists of nodes and edges. Each node represents a decision or a test on a feature.
    d. It follows a top-down approach, where the tree is built from the root node downwards. At each level, the algorithm selects the feature that best separates the data, and creates a node for that feature.
    e. The splitting criterion is a measure of how well a feature splits the data into distinct classes. Gini impurity and entropy are examples of splitting criteria.
    f. It has a special kind of decision surface: hyperrectangles
    g. It is sensitivity to axis orientation
2. (8 points) Which of the following statements are true? (select all that apply)
    a. Ensemble learning is a machine learning technique that combines multiple models to improve the accuracy of predictions
    b. The ensemble can be a strong learner even if each classifier is a weak learner, provided there are a sufficient number of weak learners in the ensemble, and they are sufficiently diverse
    c. Random forest is an ensemble learning method that combines multiple decision trees to make a prediction
    d. To improve the accuracy of the predictions, random forest tries to minimize the correlation between the trees in the ensemble, by bagging, pasting, random subspaces, random patches, etc.

(Level 2: manual exercise)

3. (18 points) Discretize the numeric attributes and splitting the dataset
    a. To split the following numeric attributes into two groups using 3 possible thresholds



    b. Please calculate the Gini impurities of 3 scenarios
    c. Which threshold will you choose?

(Level 3: computer-based exercise)

4. (60 points) Decision tree vs. SVM
    a. Use HW3-4.ipynb as the template.
    b. Use the popular scikit-learn package to implement Decision Tree (with max_depth=3 and max_depth=4)
    c. Use the popular scikit-learn package to implement Random Forest (with 30 tress, max_dpeth-3)
    d. Create two datasets with samples
        i. Dataset #1: same as the one in HW2
        ii. Dataset #2:
            1. $(-10 < x_1, x_2 < 10)$
            2. Samples with $-5 < x_1, x_2 < 5$ are positive; otherwise, are negative
        iii. Create 400 unique samples (200 are positive and 200 are negative)
        iv. Evenly split the 400 samples into two sets: one is for training and the other for testing.
    e. First, using dataset #1, train SVM (with linear kernel, C=1), decision tree (max_depth=3), decision tree (max_depth=4), and random forest (with 30 trees, max_depth=3), and test the accuracy of the unseen testing samples. Repeat 30 times and report the average accuracy and standard deviation.
    f. Second, using dataset #2, train SVM (with linear kernel, C=1), decision tree (max_depth=3), decision tree (max_depth=4), and random forest (with 30 trees, max_depth=3), and test the accuracy of the unseen testing samples. Repeat 30 times and report the average accuracy and standard deviation.
    g. Compare SVM with Decision Tree on dataset 1 and dataset 2. What do you observe? Can you explain why?
    h. Compare Decision Tree with Random Forest on dataset 1 and dataset 2. What do you observe? Can you explain why?
        i. If necessary, please use the following website to test the significant difference https://www.graphpad.com/quickcalcs/ttest1/?format=SD