HW #1 Due: 3/14/2022

1. In the lecture, we mentioned that different algorithms have different problems. Use your own words to explain the shortcomings of each of the following methods:
   - Neural networks (particularly CNN)
   - C4.5 decision tree
   - Adaboost
2. Suppose that we want to use a machine learning method to predict the rent of a house in a city. The inputs to the model include the size of the house, the built year, attached utilities, etc., and the model output is the monthly rent.
   - Suppose that a supervised-learning model is to be used. Based on the above description, between a classification model and a regression model, which one is more suitable? Explain.
   - Can the problem also be effectively solved by an unsupervised-learning algorithm?
3. Consult any statistics textbook to find the closed form of the linear regression problem given in the lecture notes, i.e., find equations for a and b to minimize

$$J = \sum_{i=1}^{10} \left(y_i - (ax_i + b)\right)^2.$$

   given $(x_i, y_i), 1 \leq i \leq N$.
4. UC Irvine has a large repository for various kinds of data. In this problem, you are asked to use the iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris) to perform the experiments. Use the k-NN classifier for the classification task with k = 7. To begin one trial, randomly draw 70% of the instances for training and the rest for testing. Repeat the trials 10 times and compute the average accuracy. Note: you can directly import iris dataset by using sklearn without downloading from the UC Irvine repository.
5. Repeat problem 6, but use 60% of the data as the training set, 20% as the validation set, and the rest 20% as test set. Vary k from 3 to 11 and use the validation set to determine the best value of k. The value of k must be determined based on average of 10 trials. Once k is known, find the average accuracy of 10 trials based on the best K.