HW #1 Due: 3/19/2021

1. Suppose that we want to use a machine learning method to predict the rent of a house in a city. The input to the model includes the size of the house, the built year, attached utilities, etc., and the model output is the monthly rent.
   a. 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?
   b. Can the problem also be effectively solved by an unsupervised-learning algorithm?
2. We skipped the AIDS detection problem in the Bayesian decision theory PPT. Now, use the Bayes theorem to confirm the answer given on the PPT file. For this problem, you need to distinguish two different conditions:
   - False positive is a conditional probability *P*(reagent is negative | patient is infected). Same argument for false negative.
   - When a patient is given a positive test result, it is actually *P*(patient is infected | reagent is positive)
3. Prove that a rectangle decision function has a VC dimension of 4 by enumerating all possible sample distributions. You can reasonably assume no two (or more) samples are on the same vertical or horizontal line.
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 $K = 5$. To begin one trial, randomly draw 70% of the samples for training and the rest for testing. Repeat the trials 10 times and compute the average accuracy. If you are familiar with the Python language, you may use the scikit-learn library to reduce the programming burden.
5. We know that *K* is a hyper-parameter in the *K*-NN algorithm. Modify your program on problem 4 to automatically select the best *K* for $1 \leq K \leq 11$.