HW #1 Due: 3/14/2022

1. Can the ChatGPT pass the Turing test? Why or why not?
2. Suppose that you want to use a machine learning method to predict the salary of a college graduate. The inputs to the model include the university name, the studied majors, years of working experiences, etc., and the model output is the monthly salary.
   - Are you going to use supervised-learning algorithms or unsupervised-learning algorithms to perform the prediction? Why?
   - 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.
3. In a binary classification problem in R^2, class "+" has samples of $\begin{bmatrix}1\\1\end{bmatrix}$, $\begin{bmatrix}2\\3\end{bmatrix}$, $\begin{bmatrix}3\\4\end{bmatrix}$, and class "−" has samples of $\begin{bmatrix}3\\1\end{bmatrix}$, $\begin{bmatrix}4\\2\end{bmatrix}$, $\begin{bmatrix}4\\3\end{bmatrix}$. We are going to use a linear classifier, represented as $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, to preform prediction. To simplify the problem, let the decision boundary passes through $\begin{bmatrix}2\\1\end{bmatrix}$ and $\begin{bmatrix}4\\4\end{bmatrix}$.
   - Find $\mathbf{w}$ and b with elements of $\mathbf{w}$ to be integers closest to zero.
   - Determine the class of the test sample $\mathbf{x} = \begin{bmatrix}3\\2\end{bmatrix}$ based on $f(\mathbf{x})$.
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 samples 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 4, but use 60% of the data as the training set, 20% as the validation set, and the rest 20% as the 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 an average of 10 trials. Then, find the average accuracy of 10 trials based on the best k.