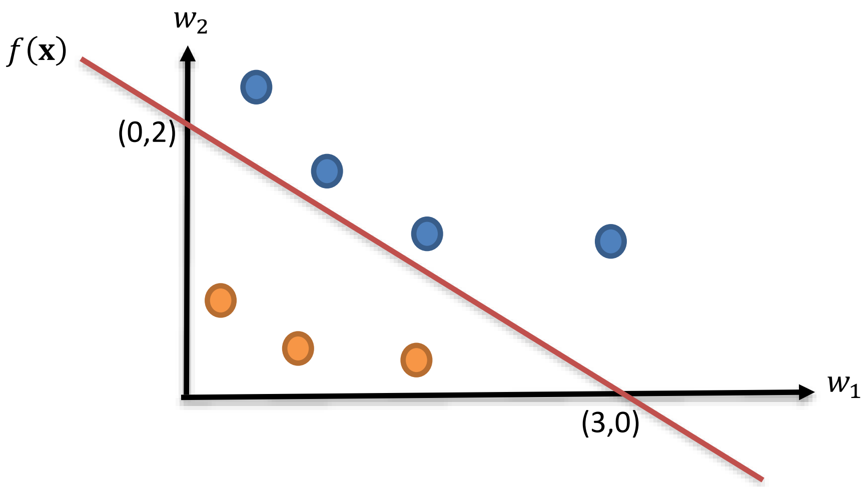
HW #1 Due: 3/05/2024

1. 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?
2. 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

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

   given $(x_i, y_i), 1 \le i \le N$.
3. In a binary classification problem in R^2, the model is represented as $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, given below.



   - Find $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ and b with elements of $\mathbf{w}$ to be positive integers closest to zero.
   - Determine the class (orange or blue) of the test sample $\mathbf{x} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$ based on $f(\mathbf{x})$.
   - If we want to make the model better, should we move the line toward upper-right or lower-left direction? Why?
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.