1. Linear Regression [15 points]
Recall the equation for ordinary least squares solution presented in class:

\[ w_{OLS} = (X^T X)^{-1} X^T y \]

Given a one dimension input, you are looking for \( y(x) = ax+b \) that achieves the minimum mean square error. Derive the formulas for \( a \) and \( b \).

2. Regularized Linear Regression [14 points]
The loss function to be optimized under ridge regression is given by:

\[ J(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T x_i))^2 + \lambda \|w\|_2^2 \]

where

\[ \lambda \|w\|_2^2 = \lambda \sum_{i=1}^{d} (w_i)^2 \]

and \( \lambda \) is the regularization constant.

There is also an alternative regularized version of regression called LASSO, which uses the L1 norm of the parameter vector as a penalty and has the following loss function to be optimized:

\[ J(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T x_i))^2 + \lambda \|w\|_1 \]

where

\[ \lambda \|w\|_1 = \lambda \sum_{i=1}^{d} |w_i| \]

(a) Suppose that we overestimate the value of \( \lambda \) for LASSO. What would happen to the number of non-zero elements of \( w \)? [2 points]

(b) Suppose that we underestimate the value of \( \lambda \) for Ridge. What would happen to the values of \( w \)? [2 points]

(c) Since \( w \) is the solution that minimizes \( J \), it follows that \( \frac{\partial J}{\partial w_i} = 0 \). Ignore the first term of \( J \) (which corresponds to the Sum Squared Error) and calculate the partial derivatives of the penalty terms for LASSO and ridge (note that the absolute value is not differentiable at \( w_i = 0 \), so consider that \( w_i \neq 0 \)). [5 points]

(d) Comparing the two derivatives, for what values of \( w_i \) will their behaviors differ? [5 points]
3. Logistic Regression [12 points]
For a logistic regression model, we can fit the parameters via maximum likelihood. To make the calculation easier, we take the log-likelihood:

\[
l(w) = \log L(w) = \sum_{i=1}^{m} y_i \log h(x_i) + (1 - y_i) \log (1 - h(x_i))
\]

where

\[
h_w(x) = \text{sigm}(w^T x) = \frac{1}{1 + e^{-w^T x}}
\]

Consider maximization using batch gradient ascent (where we use the whole training data in order to update the weights every iteration):

(a) Write down the equation for the weight update step. [6 points]
(b) Does it always converge? What is a suitable stopping condition? [6 points]

4. K-nearest Neighbors [22 points]
For this exercise you will be using a synthetic 2D dataset for binary classification. It contains separate files for training and test, with columns \(x_1, x_2\) and label (0 or 1).

(a) Run a KNN classifier with euclidean distance on the training data and plot the decision boundary (or prediction surface) for \(k=1\) and \(k=9\). Which one is more robust to outliers? [12 points]
(b) Compute and plot the accuracies on the training and test sets for \(k = [1..12]\). [10 points]

You may use any language for your implementation. Submit only the plots.

5. Decision Trees [15 points]
from Pattern Classification, by Richard O. Duda, Peter E. Hart, and David G. Stork – chapter 8
(a) Prove that for any arbitrary tree, with possible unequal branching ratios throughout, there exists a binary tree that implements the same classification functionality. [5 points]
(b) Consider a tree with just two levels - a root node connected to B leaf nodes (\(B>=2\)). What are then upper and the lower limits on the number of levels in a functionally equivalent binary tree, as a function of \(B\)? [5 points]
(c) As in b), what are the upper and lower limits on number of nodes in a functionally equivalent binary tree? [5 points]

6. Naive Bayes [22 points]
In this exercise, you will use a Naive Bayes classifier to predict whether income exceeds $50K/year based on census data from the dataset known as “Census Income” (http://archive.ics.uci.edu/ml/datasets/Adult). You should use the training data to build the model and
then measure the error when predicting on the test set (there are separate files for training and test).
Notice that the dataset has missing values and a mixture of categorical and continuous features, so you
will have to handle those conditions.

You can use any language for your implementation. Submit a description of your solution, the error on
the test set and the code. The code should contain appropriate comments to facilitate understanding.