Assignment 2

Policy: **The assignment has to be submitted in writing**. Write neatly to avoid confusion during evaluation that may invite additional clarification from students. Write your name and roll number at the top of the assignment. ***If you do not write your name and roll number you lose 5 points.***

Late submission policy: You should submit by **30-3-2021 by 23:55 PM**.
The penalty of 5% will be given per day for submission after the deadline.

Collaboration Policy: You are to complete this assignment individually. However, you are encouraged to discuss the general algorithms and ideas in the class in order to help each other answer homework questions.You are also welcome to give each other examples that are not on the assignment in order to demonstrate how to solve problems. But we require you to:

- not explicitly tell each other the answers
- not to copy answers
- not to allow your answers to be copied

In those cases where you work with one or more other people on the general discussion of the assignment and surrounding topics, we ask that you specifically record on the assignment the names of the people you were in discussion with (or "none" if you did not talk with anyone else). **This is worth five points: for each problem, your solution should either contain the names of people you talked to about it, or "none." If you do not give references for each problem, you will lose five points.** This will help resolve the situation where a mistake in general discussion led to a replicated weird error among multiple solutions. This policy has been established in order to be fair to everyone in the class.

1. For the data set given below, do the following (**10 points**)

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

a. Obtain the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
b. Obtain the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

2. Consider the XOR problem where there are four training points: (**5 points**)
(1, 1,−), (1, 0, +), (0, 1, +), (0, 0,−).
Transform the data into the following feature space:
Φ = (1, $\sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2$).
Find the maximum margin linear decision boundary in the transformed space. Show the detailed steps to obtain the linear decision boundary.

3. Derive the dual lagrangian for the linear SVM with nonseparable case with objective function given below (**5 points**)

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)^2.$$

4.Consider the one-dimensional data set shown (**10 Points**)

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

a. Classify the data point x = 4.2 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).
b. Classify the data point x = 4.2 according to its 1-, 3-, 5-, and 9-nearest neighbors using distance-weighted voting approach. Please explore on distance-weighted voting and mention how you have obtained the class label for the data point explicitly.