UCLA CS 145 Homework #2

DUE DATE: Tuesday 10/30/2018 11:59 pm

Note:

- You are expected to submit both a report and code. The submission format is specified on CCLE under HW2 description.
- "`########## Please Fill Missing Lines Here ##########`" is used where input from you is needed in the code file.

1. Decision Tree

1.1. Construct a decision tree for samples from the congressional voting records dataset with the first three attributes in the UCI machine learning repository. Information gain is used to select the attributes. Please write down the major steps in the construction process, i.e., you need to show the information gain for each candidate attribute when a new node is created in the tree.

Class	Vote for handicapped-infants?	Vote for water-project-cost-sharing	Vote for budget-resolution-adoption
Democrat	Y	N	Y
Republican	N	Y	N
Democrat	Y	Y	Y
Republican	N	Y	N
Democrat	N	Y	N
Democrat	N	Y	Y
Democrat	Y	N	Y
Democrat	Y	Y	Y
Republican	N	Y	Y
Republican	Y	Y	N
Democrat	N	N	Y
Republican	N	Y	N
Republican	N	N	N
Democrat	N	N	Y

Republican	N	N	N
Republican	N	Y	N
Democrat	Y	N	Y
Democrat	Y	N	Y
Republican	N	N	N
Republican	Y	N	Y

1.2 In DecisionTree\DecisionTree.py, fill in the missing lines for building a decision tree model, using (a) information gain and (b) gain ratio. Output the accuracy on the test data and compare the two versions of decision tree. Which attribute selection measure do you want to choose for the dataset and why.

2. Support Vector Machine

2.1 The table shown below contains 20 data points and their class labels.

Point #	x1	x2	Class (y)
1	0.52	-1	1
2	0.91	0.32	1
3	-1.48	1.23	1
4	0.01	1.44	1
5	-0.46	-0.37	1
6	0.41	2.04	1
7	0.53	0.77	1
8	-1.21	-1.1	1

9	-0.39	0.96	1
10	-0.96	0.08	1
11	2.46	2.59	-1
12	3.05	2.87	-1
13	2.2	3.04	-1
14	1.89	2.64	-1
15	4.51	-0.52	-1
16	3.06	1.3	-1
17	3.16	-0.56	-1
18	2.05	1.54	-1
19	2.34	0.72	-1
20	2.94	0.13	-1

Suppose by solving the dual form of the quadratic programming of svm, we can derive the α_i's for each data point as follows:

α_2= 0.5084

α_6= 0.4625

α_{18} = 0.9709

Others = 0

(a) Please point out the support vectors in the training points.

(b) Calculate the normal vector of the hyperplane: w

(c) Calculate the bias b, according to $b = \Sigma_{k:\alpha_k \neq 0} (y_k - w^T x_k)/N_k$, where $x_k = (x_{k1}, x_{k2})^T$ indicates the support vectors and N_k is the total number of support vectors.

(d) Write down the learned decision boundary function $f(x) = w^T x + b$ (the hyperplane)

by substituting w and b with learned values in the formula.

(e) Suppose there is a new data point $x = (-1,2)$, please use the decision boundary to predict its class label.

(f) Show a plot of the data points and your decision boundary line (x1 feature on x-axis, x2 feature on y-axis) in your report. Plot both data points and decision boundary in the same graph, and use different colors to represent points in different classes (y).

2.2 In SVM\svm.py, fill in the missing lines for support vector machines. Output the accuracy on the test data and compare (a) hard margin and soft margin SVM for linear classifier; and (b) different kernels for soft margin SVM. Which model do you want to choose?