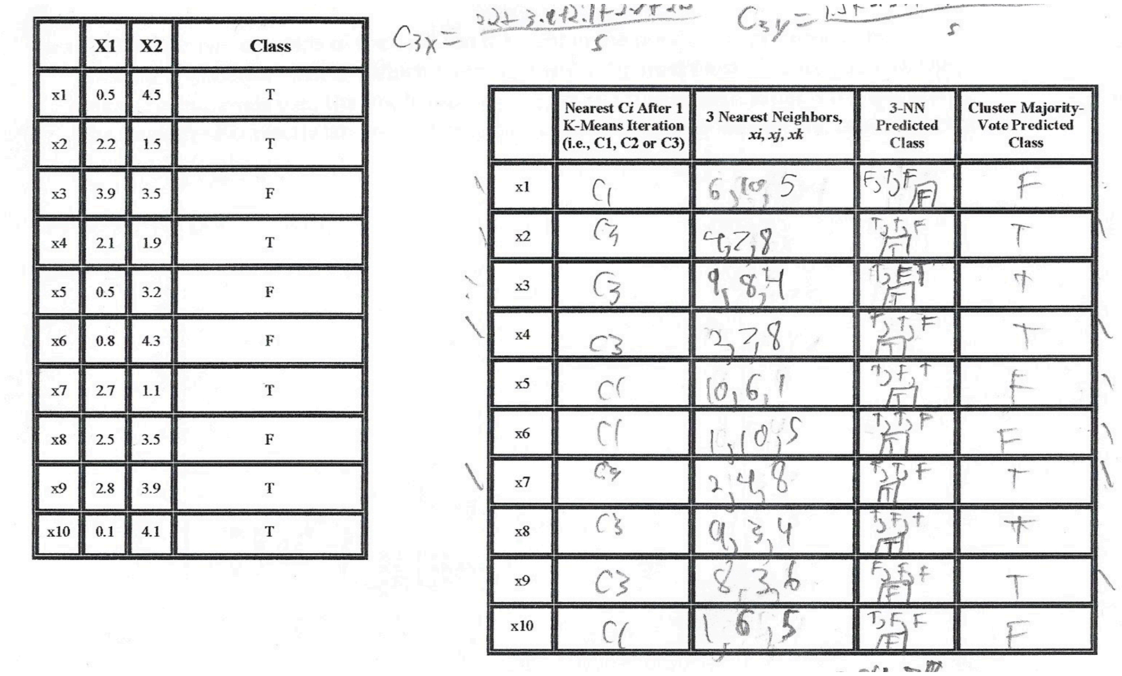
HW3 Solution Guide

P1



$C_{3x} = \frac{2.2+3.9+2.1+2.7+2.8}{5}$ $C_{3y} = \frac{1.5+}{5}$

	X1	X2	Class
x1	0.5	4.5	T
x2	2.2	1.5	T
x3	3.9	3.5	F
x4	2.1	1.9	T
x5	0.5	3.2	F
x6	0.8	4.3	F
x7	2.7	1.1	T
x8	2.5	3.5	F
x9	2.8	3.9	T
x10	0.1	4.1	T

	Nearest *Ci* After 1 K-Means Iteration (i.e., C1, C2 or C3)	3 Nearest Neighbors, *xi, xj, xk*	3-NN Predicted Class	Cluster Majority-Vote Predicted Class
x1	C1	6, 10, 5	F, T, F $\boxed{F}$	F
x2	C3	4, 7, 8	T, T, F $\boxed{T}$	T
x3	C3	9, 8, 4	T, F, T $\boxed{T}$	T
x4	C3	3, 7, 8	T, T, F $\boxed{T}$	T
x5	C1	10, 6, 1	T, F, T $\boxed{T}$	F
x6	C1	1, 10, 5	T, T, F $\boxed{T}$	F
x7	C3	2, 4, 8	T, T, F $\boxed{T}$	T
x8	C3	9, 3, 4	T, F, T $\boxed{T}$	T
x9	C3	8, 3, 6	F, F, F $\boxed{F}$	T
x10	C1	1, 6, 5	T, F, F $\boxed{F}$	F

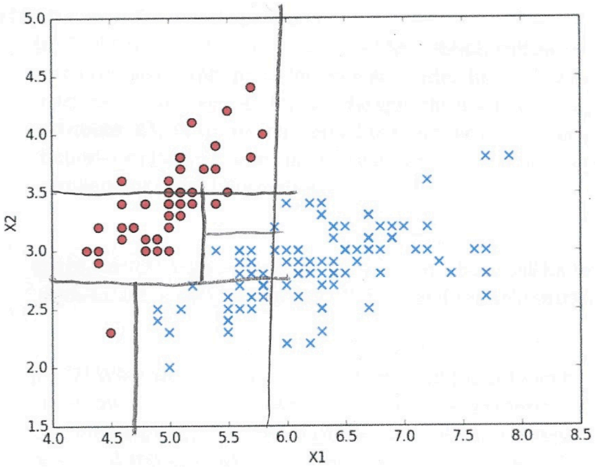
[2] a.ii (First Column), [4] b (Second Column), [2] c (Third Column), [2] d (Fourth Column)
Grading: -.5 points per incorrect label, per-column.

[4] a.i
C1 = (.94, 4.0), C2 = (0.5, 0.5), C3 = (2.68, 2.3)
Grading: 1 point per center, 1 point for some kind of work shown (Doesn't have to be the full derivations)
[2] e
Closest Cluster assigns the most correct labels, and has an accuracy of 6/10.

P2

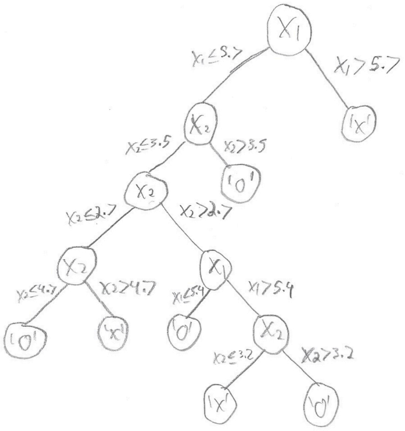
[4] a.i (Below is one possible set of splits)

Grading: -1 for slanted and not axis parallel lines. -2 for boxes with multiple class kinds.



[2] a.ii

Grading: -1 for each node not matching the DT.



[14] b.i

Grading: -2 for each incorrect IG split value.

IG X1

$$IG = .91829 - \left(\frac{4}{9}\right) - \left(\frac{5}{9}\right)(.72192) = .07278$$

IG X2

$IG =$ .07278 (Same as X1)

X3 (Threshold 2)

$$IG = .9189 - 0 - \left(\frac{8}{9}\right)(.9549) = .06991$$

X3 (Threshold 3.5)

$$IG = .9189 - (\frac{2}{9}) - \left(\frac{7}{9}\right)(.8631) = .02475$$

X3 (Threshold 7.25)

$$IG = .9189 - 0 - \left(\frac{7}{9}\right)(.59167) = .4581$$

[2] b.ii

X3 (Threshold 7.25)

$$IG = .9189 - 0 - \left(\frac{7}{9}\right)(.59167) = .4581$$

[2] b.iii

Grading: 1 pt for answer, 1 pt for explanation.

‘Instance’ should **not** be included when considering splits. It is a unique value, known to have no bearing on predicting classes. Also, it will be split on first since each unique value can be used to perfectly classify its row’s class label. Also, generalization accuracy would decrease if this attribute was considered.