Name: Wenliang Sun

3.9 For Setosa, sepal length > sepal width > petal length > petal width. For Versicolour and Virginiica, sepal length > sepal width and petal length > petal width, but although sepal length > petal length, petal length > sepal width.

3.11 We would expect such a distribution if the three species of Iris can be ordered according to their size, and if petal length and width are both correlated to the size of the plant and each other.

4.2(a) $\text{Gini}=1-2\times0.52=0.5$

(c) The gini for Male is $1 - 2 \times 0.52 = 0.5$. The gini for Female is also 0.5. Therefore, the overall gini for Gender is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

4.5(b)

The overall gini before splitting is:

$$G_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$$

The gain in gini after splitting on A is:

$$\begin{aligned} G_{A=T} &= 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898 \\ G_{A=F} &= 1 = \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0 \\ \Delta &= G_{orig} - 7/10 G_{A=T} - 3/10 G_{A=F} = 0.1371 \end{aligned}$$

The gain in gini after splitting on B is:

$$\begin{aligned} G_{B=T} &= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750 \\ G_{B=F} &= 1 = \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778 \\ \Delta &= G_{orig} - 4/10 G_{B=T} - 6/10 G_{B=F} = 0.1633 \end{aligned}$$

Therefore, attribute B will be chosen to split the node.

(c) Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

4.7(a)

The error rate for the data without partitioning on any attribute is

$$E_{orig} = 1 - \max(\frac{50}{100}, \frac{50}{100}) = \frac{50}{100}.$$

After splitting on attribute A, the gain in error rate is:

	$A = T$	$A = F$
+	25	25
−	0	50

$$E_{A=T} = 1 - \max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0$$

$$E_{A=F} = 1 - \max(\frac{25}{75}, \frac{50}{75}) = \frac{25}{75}$$

$$\Delta_A = E_{orig} - \frac{25}{100}E_{A=T} - \frac{75}{100}E_{A=F} = \frac{25}{100}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
+	30	20
−	20	30

$$E_{B=T} = \frac{20}{50}$$

$$E_{B=F} = \frac{20}{50}$$

$$\Delta_B = E_{orig} - \frac{50}{100}E_{B=T} - \frac{50}{100}E_{B=F} = \frac{10}{100}$$

After splitting on attribute C, the gain in error rate is:

	$C = T$	$C = F$
+	25	25
−	25	25

$$E_{C=T} = \frac{25}{50}$$

$$E_{C=F} = \frac{25}{50}$$

$$\Delta_C = E_{orig} - \frac{50}{100}E_{C=T} - \frac{50}{100}E_{C=F} = \frac{0}{100} = 0$$

The algorithm chooses attribute A because it has the highest gain.

(b)

Because the $A = T$ child node is pure, no further splitting is needed. For the $A = F$ child node, the distribution of training instances is:

B	C	Class label	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The classification error of the $A = F$ child node is:

$$E_{orig} = \frac{25}{75}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
$+$	25	0
$-$	20	30

$$E_{B=T} = \frac{20}{45}$$
$$E_{B=F} = 0$$
$$\Delta_B = E_{orig} - \frac{45}{75}E_{B=T} - \frac{20}{75}E_{B=F} = \frac{5}{75}$$

After splitting on attribute C, the gain in error rate is:

	$C = T$	$C = F$
$+$	0	25
$-$	25	25

$$E_{C=T} = \frac{0}{25}$$
$$E_{C=F} = \frac{25}{50}$$
$$\Delta_C = E_{orig} - \frac{25}{75}E_{C=T} - \frac{50}{75}E_{C=F} = 0$$

The split will be made on attribute B.

(c) 20 instances are misclassified. (The error rate is 20 .)

5.4(a) The accuracies of the rules are 80% (for R1), 75% (for R2), and 52.6% (for R3), respectively. Therefore R1 is the best candidate and R3 is the worst candidate according to rule accuracy.

(d) The Laplace measure of the rules are 71.43% (for R1), 73.81% (for R2), and 52.6% (for R3), respectively. Therefore R2 is the best candidate and R3 is the worst candidate according to the Laplace measure.