CS4445 A Term 2008
Homework 3 Solutions

Contents

1 Linear Regression

See text for the descriptions of the algorithms and procedures.

$$
\begin{aligned}
\textbf{FIQ} &= \\
+17.074 &* \textbf{CCMIDSA} \\
+7.436 &* \textbf{male} \\
-0.109 &* \textbf{TOTVOL} \\
+0.071 &* \textbf{WEIGHT} \\
+99.614 &
\end{aligned}
$$

2 Regression Trees and Model Trees

2.1 Gender Attribute Conversion

Let us start by converting the nominal **GENDER** attribute into several numeric boolean attributes. We have, to begin with, the instances in this form:

#	CCMIDSA	GENDER	TOTVOL	WEIGHT	FIQ
1	*6.08*	*female*	*1005*	*57.607*	*96*
2	*5.73*	*female*	*963*	*58.968*	*89*
3	*7.99*	*female*	*1281*	*63.958*	*101*
4	*8.42*	*female*	*1272*	*61.69*	*103*
5	*6.84*	*female*	*1079*	*107.503*	*96*
6	*6.43*	*female*	*1070*	*83.009*	*126*
7	*7.6*	*male*	*1347*	*97.524*	*94*
8	*6.03*	*male*	*1029*	*81.648*	*97*
9	*7.52*	*male*	*1204*	*79.38*	*113*
10	*7.67*	*male*	*1160*	*72.576*	*124*

We proceed by converting the **GENDER** attribute to binary values. This attribute has the following values: *male*,*female*

2.1.1 GENDER=*male*

#	CCMIDSA	GENDER	TOTVOL	WEIGHT	FIQ
7	*7.6*	*male*	*1347*	*97.524*	*94*
8	*6.03*	*male*	*1029*	*81.648*	*97*
9	*7.52*	*male*	*1204*	*79.38*	*113*
10	*7.67*	*male*	*1160*	*72.576*	*124*

We see the average class over these instances is 107.000.

2.1.2 GENDER=*female*

#	CCMIDSA	GENDER	TOTVOL	WEIGHT	FIQ
1	*6.08*	*female*	*1005*	*57.607*	*96*
2	*5.73*	*female*	*963*	*58.968*	*89*
3	*7.99*	*female*	*1281*	*63.958*	*101*
4	*8.42*	*female*	*1272*	*61.69*	*103*
5	*6.84*	*female*	*1079*	*107.503*	*96*
6	*6.43*	*female*	*1070*	*83.009*	*126*

We see the average class over these instances is 101.833.

2.1.3 New Attributes

The attribute values sorted by average class value are thus: *male*, *female*. We create new binary attributes accordingly:

- *male*

2.1.4 Converted Dataset

Our final dataset, in its converted form, is:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.08*	*0*	*1005*	*57.607*	*96*
2	*5.73*	*0*	*963*	*58.968*	*89*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
5	*6.84*	*0*	*1079*	*107.503*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
8	*6.03*	*1*	*1029*	*81.648*	*97*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

2.2 Tree Construction

Next we will construct a model tree. First we compute the standard deviation over all the instances arriving with 12.153. We will therefore use $0.05 * 12.153 = 0.608$ as one of our stopping criteria. We will also stop splitting a node if it contains less than 4 instances. Our tree so far:

?

The instances applicable to this node are:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.08*	*0*	*1005*	*57.607*	*96*
2	*5.73*	*0*	*963*	*58.968*	*89*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
5	*6.84*	*0*	*1079*	*107.503*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
8	*6.03*	*1*	*1029*	*81.648*	*97*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

Let us first check stop conditions. The standard deviation over these instances is 12.153 and there are/is 10 of them. The conditions for stopping are not met so we will look for the best split point.

2.2.1 Split

2.2.1.1 Split Points for CCMIDSA

Sorting our instances by **CCMIDSA** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
2	*5.73*	*0*	*963*	*58.968*	*89*
8	*6.03*	*1*	*1029*	*81.648*	*97*
1	*6.08*	*0*	*1005*	*57.607*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
5	*6.84*	*0*	*1079*	*107.503*	*96*
9	*7.52*	*1*	*1204*	*79.38*	*113*
7	*7.6*	*1*	*1347*	*97.524*	*94*
10	*7.67*	*1*	*1160*	*72.576*	*124*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*

This results in these split points: *5.880*, *6.055*, *6.255*, *6.635*, *7.180*, *7.560*, *7.635*, *7.830*, *8.205*.

2.2.1.2 Split Points for male

Sorting our instances by **male** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.08*	*0*	*1005*	*57.607*	*96*
2	*5.73*	*0*	*963*	*58.968*	*89*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
5	*6.84*	*0*	*1079*	*107.503*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
8	*6.03*	*1*	*1029*	*81.648*	*97*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

This results in these split points: *0.500*.

2.2.1.3 Split Points for TOTVOL

Sorting our instances by **TOTVOL** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
2	*5.73*	*0*	*963*	*58.968*	*89*
1	*6.08*	*0*	*1005*	*57.607*	*96*
8	*6.03*	*1*	*1029*	*81.648*	*97*
6	*6.43*	*0*	*1070*	*83.009*	*126*
5	*6.84*	*0*	*1079*	*107.503*	*96*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*
7	*7.6*	*1*	*1347*	*97.524*	*94*

This results in these split points: *984.000*, *1017.000*, *1049.500*, *1074.500*, *1119.500*, *1182.000*, *1238.000*, *1276.500*, *1314.000*.

2.2.1.4 Split Points for WEIGHT

Sorting our instances by **WEIGHT** gives us:

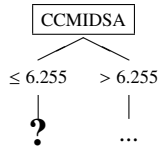
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.08*	*0*	*1005*	*57.607*	*96*
2	*5.73*	*0*	*963*	*58.968*	*89*
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
8	*6.03*	*1*	*1029*	*81.648*	*97*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
5	*6.84*	*0*	*1079*	*107.503*	*96*

This results in these split points: *58.288*, *60.329*, *62.824*, *68.267*, *75.978*, *80.514*, *82.328*, *90.267*, *102.513*.

2.2.1.5 All Split Points

attribute	split	instaces ≤	instaces >	std. dev. ≤	std. dev. >	SDR
CCMIDSA	5.880	2	1 3 4 5 6 7 8 9 10	0.000	11.692	1.630
CCMIDSA	6.055	2 8	1 3 4 5 6 7 9 10	4.000	11.978	1.770
CCMIDSA	6.255	1 2 8	3 4 5 6 7 9 10	3.559	12.064	2.640
CCMIDSA	6.635	1 2 6 8	3 4 5 7 9 10	14.195	10.383	0.245
CCMIDSA	7.180	1 2 5 6 8	3 4 7 9 10	12.921	10.450	0.467
CCMIDSA	7.560	1 2 5 6 8 9	3 4 7 10	12.641	11.192	0.091
CCMIDSA	7.635	1 2 5 6 7 8 9	3 4 10	12.105	10.403	0.558
CCMIDSA	7.830	1 2 5 6 7 8 9 10	3 4	13.536	1.000	1.124
CCMIDSA	8.205	1 2 3 5 6 7 8 9 10	4	12.806	0.000	0.627
male	0.500	1 2 3 4 5 6	7 8 9 10	11.682	12.186	0.269
TOTVOL	984.000	2	1 3 4 5 6 7 8 9 10	0.000	11.692	1.630
TOTVOL	1017.000	1 2	3 4 5 6 7 8 9 10	3.500	11.872	1.955
TOTVOL	1049.500	1 2 8	3 4 5 6 7 9 10	3.559	12.064	2.640
TOTVOL	1074.500	1 2 6 8	3 4 5 7 9 10	14.195	10.383	0.245
TOTVOL	1119.500	1 2 5 6 8	3 4 7 9 10	12.921	10.450	0.467
TOTVOL	1182.000	1 2 5 6 8 10	3 4 7 9	14.625	6.796	0.659
TOTVOL	1238.000	1 2 5 6 8 9 10	3 4 7	13.851	3.859	1.300
TOTVOL	1276.500	1 2 4 5 6 8 9 10	3 7	12.990	3.500	1.060
TOTVOL	1314.000	1 2 3 4 5 6 8 9 10	7	12.329	0.000	1.057
WEIGHT	58.288	1	2 3 4 5 6 7 8 9 10	0.000	12.506	0.898
WEIGHT	60.329	1 2	3 4 5 6 7 8 9 10	3.500	11.872	1.955
WEIGHT	62.824	1 2 4	3 5 6 7 8 9 10	5.715	12.601	1.618
WEIGHT	68.267	1 2 3 4	5 6 7 8 9 10	5.403	13.325	1.997
WEIGHT	75.978	1 2 3 4 10	5 6 7 8 9	11.741	12.416	0.074
WEIGHT	80.514	1 2 3 4 9 10	5 6 7 8	11.397	13.179	0.043
WEIGHT	82.328	1 2 3 4 8 9 10	5 6 7	10.859	14.636	0.161
WEIGHT	90.267	1 2 3 4 6 8 9 10	5 7	12.634	1.000	1.846
WEIGHT	102.513	1 2 3 4 6 7 8 9 10	5	12.506	0.000	0.898

We see that the best split is by CCMIDSA around 6.255. We create nodes based on this split and continue. Our tree so far:



The instances applicable to this node are:

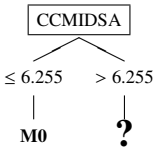
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.08*	*0*	*1005*	*57.607*	*96*
2	*5.73*	*0*	*963*	*58.968*	*89*
8	*6.03*	*1*	*1029*	*81.648*	*97*

Let us first check stop conditions. The standard deviation over these instances is 3.559 and there are/is 3 of them. The number of instances here is below our minimum of 4. We label this node with a model. The average class over these instances is 94.000. We will use this as our RegressionTree model at the node. We also run a

linearRegression fit on these instances and note the coefficients produced. The nonregressionTree model here will therefore be:

$$\begin{aligned} \textbf{FIQ} &= \\ +13.464 &* \textbf{CCMIDSA} \\ +0.298 &* \textbf{male} \\ +0.055 &* \textbf{TOTVOL} \\ +0.003 &* \textbf{WEIGHT} \\ + - 40.845 & \end{aligned}$$

Our tree so far:



The instances applicable to this node are:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
5	*6.84*	*0*	*1079*	*107.503*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

Let us first check stop conditions. The standard deviation over these instances is 12.064 and there are/is 7 of them. The conditions for stopping are not met so we will look for the best split point.

2.2.2 Split

2.2.2.1 Split Points for CCMIDSA

Sorting our instances by **CCMIDSA** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
6	*6.43*	*0*	*1070*	*83.009*	*126*
5	*6.84*	*0*	*1079*	*107.503*	*96*
9	*7.52*	*1*	*1204*	*79.38*	*113*
7	*7.6*	*1*	*1347*	*97.524*	*94*
10	*7.67*	*1*	*1160*	*72.576*	*124*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*

This results in these split points: *6.635*, *7.180*, *7.560*, *7.635*, *7.830*, *8.205*.

2.2.2.2 Split Points for male

Sorting our instances by **male** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
5	*6.84*	*0*	*1079*	*107.503*	*96*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

This results in these split points: *0.500*.

2.2.2.3 Split Points for TOTVOL

Sorting our instances by **TOTVOL** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
6	*6.43*	*0*	*1070*	*83.009*	*126*
5	*6.84*	*0*	*1079*	*107.503*	*96*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*
7	*7.6*	*1*	*1347*	*97.524*	*94*

This results in these split points: *1074.500*, *1119.500*, *1182.000*, *1238.000*, *1276.500*, *1314.000*.

2.2.2.4 Split Points for WEIGHT

Sorting our instances by **WEIGHT** gives us:

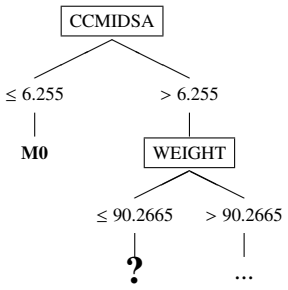
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
6	*6.43*	*0*	*1070*	*83.009*	*126*
7	*7.6*	*1*	*1347*	*97.524*	*94*
5	*6.84*	*0*	*1079*	*107.503*	*96*

This results in these split points: *62.824*, *68.267*, *75.978*, *81.195*, *90.267*, *102.513*.

2.2.2.5 All Split Points

attribute	split	instaces ≤	instaces >	std. dev. ≤	std. dev. >	SDR
CCMIDSA	6.635	6	3 4 5 7 9 10	0.000	10.383	3.165
CCMIDSA	7.180	5 6	3 4 7 9 10	15.000	10.450	0.315
CCMIDSA	7.560	5 6 9	3 4 7 10	12.284	11.192	0.405
CCMIDSA	7.635	5 6 7 9	3 4 10	13.103	10.403	0.119
CCMIDSA	7.830	5 6 7 9 10	3 4	13.500	1.000	2.136
CCMIDSA	8.205	3 5 6 7 9 10	4	12.832	0.000	1.065
male	0.500	3 4 5 6	7 9 10	11.543	12.392	0.157
TOTVOL	1074.500	6	3 4 5 7 9 10	0.000	10.383	3.165
TOTVOL	1119.500	5 6	3 4 7 9 10	15.000	10.450	0.315
TOTVOL	1182.000	5 6 10	3 4 7 9	13.695	6.796	2.312
TOTVOL	1238.000	5 6 9 10	3 4 7	11.903	3.859	3.609
TOTVOL	1276.500	4 5 6 9 10	3 7	11.638	3.500	2.752
TOTVOL	1314.000	3 4 5 6 9 10	7	11.442	0.000	2.257
WEIGHT	62.824	4	3 5 6 7 9 10	0.000	12.832	1.065
WEIGHT	68.267	3 4	5 6 7 9 10	1.000	13.500	2.136
WEIGHT	75.978	3 4 10	5 6 7 9	10.403	13.103	0.119
WEIGHT	81.195	3 4 9 10	5 6 7	9.148	14.636	0.564
WEIGHT	90.267	3 4 6 9 10	5 7	10.327	1.000	4.403
WEIGHT	102.513	3 4 6 7 9 10	5	11.880	0.000	1.881

We see that the best split is by WEIGHT around 90.267. We create nodes based on this split and continue. Our tree so far:



The instances applicable to this node are:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
6	*6.43*	*0*	*1070*	*83.009*	*126*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

Let us first check stop conditions. The standard deviation over these instances is 10.327 and there are/is 5 of them. The conditions for stopping are not met so we will look for the best split point.

2.2.3 Split

2.2.3.1 Split Points for CCMIDSA

Sorting our instances by **CCMIDSA** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
6	*6.43*	*0*	*1070*	*83.009*	*126*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*

This results in these split points: *6.975*, *7.595*, *7.830*, *8.205*.

2.2.3.2 Split Points for male

Sorting our instances by **male** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
6	*6.43*	*0*	*1070*	*83.009*	*126*
9	*7.52*	*1*	*1204*	*79.38*	*113*
10	*7.67*	*1*	*1160*	*72.576*	*124*

This results in these split points: *0.500*.

2.2.3.3 Split Points for TOTVOL

Sorting our instances by **TOTVOL** gives us:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
6	*6.43*	*0*	*1070*	*83.009*	*126*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*

This results in these split points: *1115.000*, *1182.000*, *1238.000*, *1276.500*.

2.2.3.4 Split Points for WEIGHT

Sorting our instances by **WEIGHT** gives us:

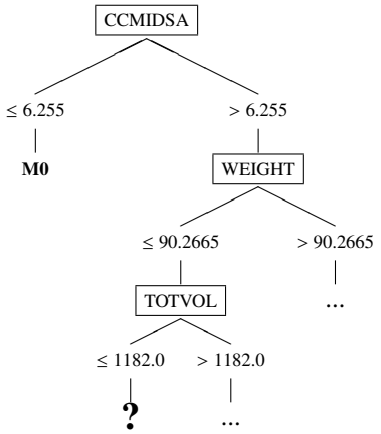
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
4	*8.42*	*0*	*1272*	*61.69*	*103*
3	*7.99*	*0*	*1281*	*63.958*	*101*
10	*7.67*	*1*	*1160*	*72.576*	*124*
9	*7.52*	*1*	*1204*	*79.38*	*113*
6	*6.43*	*0*	*1070*	*83.009*	*126*

This results in these split points: *62.824*, *68.267*, *75.978*, *81.195*.

2.2.3.5 All Split Points

attribute	split	instaces ≤	instaces >	std. dev. ≤	std. dev. >	SDR
CCMIDSA	6.975	6	3 4 9 10	0.000	9.148	3.008
CCMIDSA	7.595	6 9	3 4 10	6.500	10.403	1.485
CCMIDSA	7.830	6 9 10	3 4	5.715	1.000	6.497
CCMIDSA	8.205	3 6 9 10	4	9.975	0.000	2.347
male	0.500	3 4 6	9 10	11.343	5.500	1.321
TOTVOL	1115.000	6	3 4 9 10	0.000	9.148	3.008
TOTVOL	1182.000	6 10	3 4 9	1.000	5.249	6.777
TOTVOL	1238.000	6 9 10	3 4	5.715	1.000	6.497
TOTVOL	1276.500	4 6 9 10	3	9.233	0.000	2.940
WEIGHT	62.824	4	3 6 9 10	0.000	9.975	2.347
WEIGHT	68.267	3 4	6 9 10	1.000	5.715	6.497
WEIGHT	75.978	3 4 10	6 9	10.403	6.500	1.485
WEIGHT	81.195	3 4 9 10	6	9.148	0.000	3.008

We see that the best split is by TOTVOL around 1182.000. We create nodes based on this split and continue. Our tree so far:



The instances applicable to this node are:

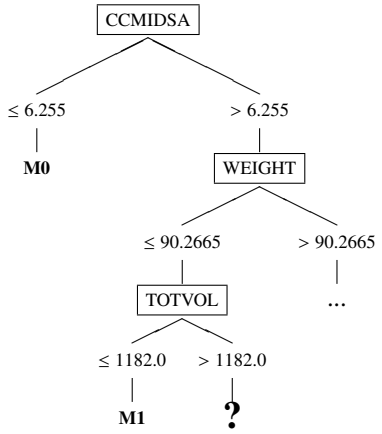
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
6	*6.43*	*0*	*1070*	*83.009*	*126*
10	*7.67*	*1*	*1160*	*72.576*	*124*

Let us first check stop conditions. The standard deviation over these instances is 1.000 and there are/is 2 of them. The number of instances here is below our minimum of 4. We label this node with a model. The average class over these instances is 125.000. We will use this as our RegressionTree model at the node. We also run a linearRegression fit on these instances and note the coefficients produced. The non-

regressionTree model here will therefore be:

$$\begin{aligned} \textbf{FIQ} &= \\ -0.403 &* \textbf{CCMIDSA} \\ -0.500 &* \textbf{male} \\ -0.006 &* \textbf{TOTVOL} \\ +0.048 &* \textbf{WEIGHT} \\ +130.559 & \end{aligned}$$

Our tree so far:



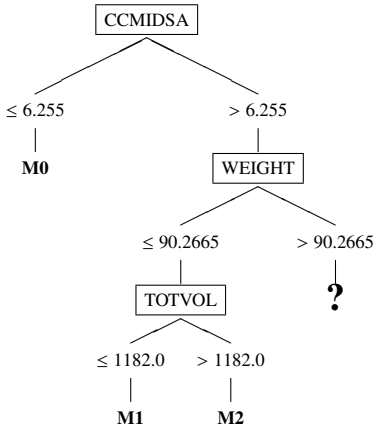
The instances applicable to this node are:

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*0*	*1281*	*63.958*	*101*
4	*8.42*	*0*	*1272*	*61.69*	*103*
9	*7.52*	*1*	*1204*	*79.38*	*113*

Let us first check stop conditions. The standard deviation over these instances is 5.249 and there are/is 3 of them. The number of instances here is below our minimum of 4. We label this node with a model. The average class over these instances is 105.667. We will use this as our RegressionTree model at the node. We also run a linearRegression fit on these instances and note the coefficients produced. The non-regressionTree model here will therefore be:

$$\begin{aligned} \textbf{FIQ} &= \\ +3.786 &* \textbf{CCMIDSA} \\ +4.623 &* \textbf{male} \\ -0.085 &* \textbf{TOTVOL} \\ +0.172 &* \textbf{WEIGHT} \\ +168.092 & \end{aligned}$$

Our tree so far:



The instances applicable to this node are:

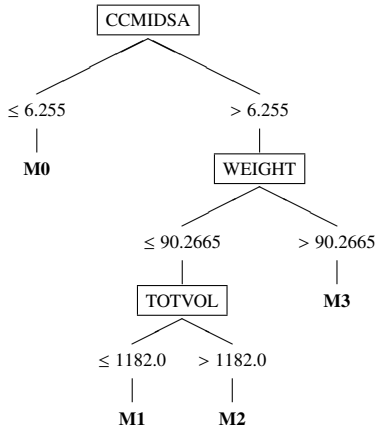
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
5	*6.84*	*0*	*1079*	*107.503*	*96*
7	*7.6*	*1*	*1347*	*97.524*	*94*

Let us first check stop conditions. The standard deviation over these instances is 1.000 and there are/is 2 of them. The number of instances here is below our minimum of 4. We label this node with a model. The average class over these instances is 95.000. We will use this as our RegressionTree model at the node. We also run a linearRegression fit on these instances and note the coefficients produced. The non-regressionTree model here will therefore be:

$$
\begin{aligned}
\textbf{FIQ} &= \\
-0.658 &* \textbf{CCMIDSA} \\
-0.500 &* \textbf{male} \\
-0.002 &* \textbf{TOTVOL} \\
+0.050 &* \textbf{WEIGHT} \\
+97.127 &
\end{aligned}
$$

2.2.4 Final Tree

Our final tree now is:



2.2.5 Final Models

The various sub models produced are summarized below:

2.2.6 M0

M0 for RegressionTree

$$\textbf{FIQ} = \textbf{94.000}$$

M0 for ModelTree

$$\begin{aligned} \textbf{FIQ} &= \\ +13.464 &* \textbf{CCMIDSA} \\ +0.298 &* \textbf{male} \\ +0.055 &* \textbf{TOTVOL} \\ +0.003 &* \textbf{WEIGHT} \\ +-40.845 & \end{aligned}$$

2.2.7 M1

M1 for RegressionTree

$$\textbf{FIQ} = \textbf{125.000}$$

M1 for ModelTree

FIQ	=	
−0.403	∗	**CCMIDSA**
−0.500	∗	**male**
−0.006	∗	**TOTVOL**
+0.048	∗	**WEIGHT**
+130.559		

2.2.8 M2

M2 for RegressionTree

FIQ = 105.667

M2 for ModelTree

FIQ	=	
+3.786	∗	**CCMIDSA**
+4.623	∗	**male**
−0.085	∗	**TOTVOL**
+0.172	∗	**WEIGHT**
+168.092		

2.2.9 M3

M3 for RegressionTree

FIQ = 95.000

M3 for ModelTree

$$
\begin{aligned}
\textbf{FIQ} &= \\
-0.658 &* \textbf{CCMIDSA} \\
-0.500 &* \textbf{male} \\
-0.002 &* \textbf{TOTVOL} \\
+0.050 &* \textbf{WEIGHT} \\
+97.127 &
\end{aligned}
$$

3 Testing

We test our three models on four sample instances. We have the following instances to evaluate:

#	CCMIDSA	GENDER	TOTVOL	WEIGHT	FIQ
1	*6.22*	*female*	*1035*	*64.184*	*87*
2	*6.48*	*female*	*1034*	*62.143*	*127*
3	*7.99*	*male*	*1173*	*61.236*	*101*
4	*6.59*	*male*	*1100*	*88.452*	*114*

We begin by converting them to our format (with GENDER binarized):

#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.22*	*0*	*1035*	*64.184*	*87*
2	*6.48*	*0*	*1034*	*62.143*	*127*
3	*7.99*	*1*	*1173*	*61.236*	*101*
4	*6.59*	*1*	*1100*	*88.452*	*114*

We can now use our three methods to predict the target value for each instance.

3.1 Instance 1

Let us classify:

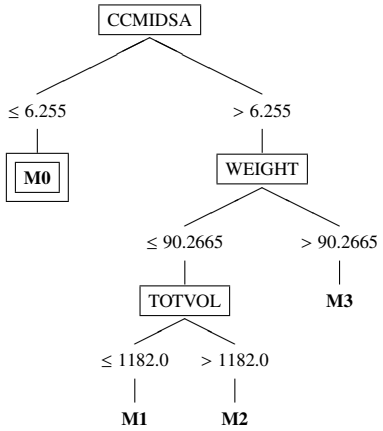
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
1	*6.22*	*0*	*1035*	*64.184*	*87*

3.1.1 Linear Regression

FIQ	=		
+17.074	∗	6.220	*[CCMIDSA]*
+7.436	∗	0.000	*[male]*
−0.109	∗	1035.000	*[TOTVOL]*
+0.071	∗	64.184	*[WEIGHT]*
+99.614	=	97.745	

3.1.2 Model Tree

We see that this instance trickled down to this (marked) model in our tree:



We will use this for our ModelTree and RegressionTree predictions.

FIQ	=		
+13.464	∗	6.220	*[CCMIDSA]*
+0.298	∗	0.000	*[male]*
+0.055	∗	1035.000	*[TOTVOL]*
+0.003	∗	64.184	*[WEIGHT]*
−40.845	=	99.539	

3.1.3 Regression Tree

FIQ = 94.000

3.2 Instance 2

Let us classify:

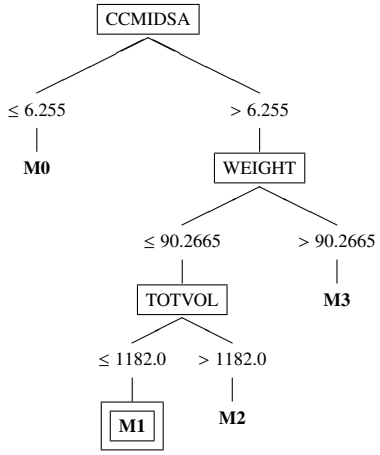
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
2	*6.48*	*0*	*1034*	*62.143*	*127*

3.2.1 Linear Regression

$$
\begin{aligned}
\textbf{FIQ} &= \\
+17.074 &* 6.480 \quad \textit{[CCMIDSA]} \\
+7.436 &* 0.000 \quad \textit{[male]} \\
-0.109 &* 1034.000 \quad \textit{[TOTVOL]} \\
+0.071 &* 62.143 \quad \textit{[WEIGHT]} \\
+99.614 &= \boxed{102.147}
\end{aligned}
$$

3.2.2 Model Tree

We see that this instance trickled down to this (marked) model in our tree:



We will use this for our ModelTree and RegressionTree predictions.

$$
\begin{aligned}
\textbf{FIQ} &= \\
-0.403 &* 6.480 \quad \textit{[CCMIDSA]} \\
-0.500 &* 0.000 \quad \textit{[male]} \\
-0.006 &* 1034.000 \quad \textit{[TOTVOL]} \\
+0.048 &* 62.143 \quad \textit{[WEIGHT]} \\
+130.559 &= \boxed{125.180}
\end{aligned}
$$

3.2.3 Regression Tree

FIQ = 125.000

3.3 Instance 3

Let us classify:

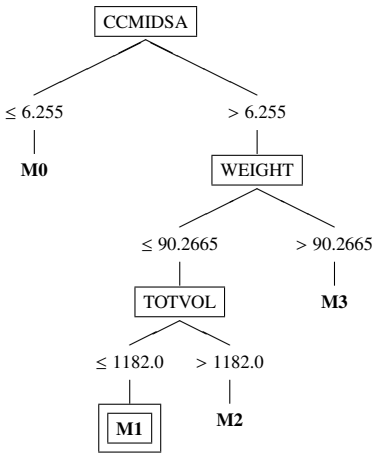
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
3	*7.99*	*1*	*1173*	*61.236*	*101*

3.3.1 Linear Regression

FIQ	=		
+17.074	∗	7.990	*[CCMIDSA]*
+7.436	∗	1.000	*[male]*
−0.109	∗	1173.000	*[TOTVOL]*
+0.071	∗	61.236	*[WEIGHT]*
+99.614	=	120.171	

3.3.2 Model Tree

We see that this instance trickled down to this (marked) model in our tree:



We will use this for our ModelTree and RegressionTree predictions.

FIQ	=		
−0.403	∗	7.990	*[CCMIDSA]*
−0.500	∗	1.000	*[male]*
−0.006	∗	1173.000	*[TOTVOL]*
+0.048	∗	61.236	*[WEIGHT]*
+130.559	=	123.255	

3.3.3 Regression Tree

FIQ = 125.000

3.4 Instance 4

Let us classify:

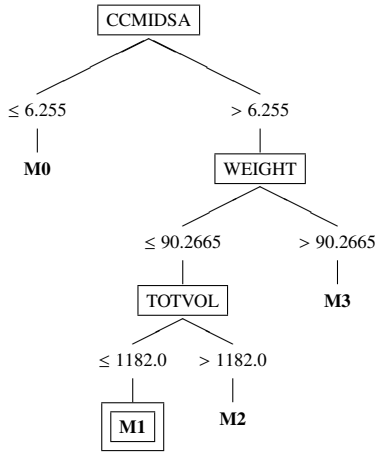
#	CCMIDSA	male	TOTVOL	WEIGHT	FIQ
4	*6.59*	*1*	*1100*	*88.452*	*114*

3.4.1 Linear Regression

FIQ	=		
+17.074	∗	6.590	*[CCMIDSA]*
+7.436	∗	1.000	*[male]*
−0.109	∗	1100.000	*[TOTVOL]*
+0.071	∗	88.452	*[WEIGHT]*
+99.614	=	106.156	

3.4.2 Model Tree

We see that this instance trickled down to this (marked) model in our tree:



We will use this for our ModelTree and RegressionTree predictions.

FIQ	=		
−0.403	∗	6.590	*[CCMIDSA]*
−0.500	∗	1.000	*[male]*
−0.006	∗	1100.000	*[TOTVOL]*
+0.048	∗	88.452	*[WEIGHT]*
+130.559	=	125.530	

3.4.3 Regression Tree

FIQ = 125.000

3.5 Testing Results

3.5.1 Predicted Values

Our three methods made the following predictions:

#	FIQ	Linear Regression	Model Tree	Regression Tree
1	*87*	*97.745*	*99.539*	*94.000*
2	*127*	*102.147*	*125.180*	*125.000*
3	*101*	*120.171*	*123.255*	*125.000*
4	*114*	*106.156*	*125.530*	*125.000*

3.5.2 Error Measures

3.5.2.1 Root-Mean Squared Error

For the root-mean squared error, we use the following formula from the class text:

$$\left(\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n}\right)^{\frac{1}{2}}$$

Linear Regression

$$\left(\frac{(97.745 - 87.000)^2 + (102.147 - 127.000)^2 + (120.171 - 101.000)^2 + (106.156 - 114.000)^2}{4}\right)^{\frac{1}{2}} = 17.045$$

Model Tree

$$\left(\frac{(99.539 - 87.000)^2 + (125.180 - 127.000)^2 + (123.255 - 101.000)^2 + (125.530 - 114.000)^2}{4}\right)^{\frac{1}{2}} = 14.043$$

Regression Tree

$$\left(\frac{(94.000 - 87.000)^2 + (125.000 - 127.000)^2 + (125.000 - 101.000)^2 + (125.000 - 114.000)^2}{4}\right)^{\frac{1}{2}} = 13.693$$

3.5.2.2 Mean Absolute Error

For the mean absolute error, we use the following formula from the class text:

$$\frac{|p_1 - a_1| + ... + |p_n - a_n|}{n}$$

Linear Regression

$$\frac{|97.745 - 87.000| + |102.147 - 127.000| + |120.171 - 101.000| + |106.156 - 114.000|}{4} = 15.653$$

Model Tree

$$\frac{|99.539 - 87.000| + |125.180 - 127.000| + |123.255 - 101.000| + |125.530 - 114.000|}{4} = 12.036$$

Regression Tree

$$\frac{|94.000 - 87.000| + |125.000 - 127.000| + |125.000 - 101.000| + |125.000 - 114.000|}{4} = 11.000$$

3.5.3 Error Summary

And we have our accuracy measures:

measure	Linear Regression	Model Tree	Regression Tree
root mean-squared error	17.045	14.043	13.693
mean absolute error	15.653	12.036	11.000