Classification Results - May 20, 2010

James Long

1 Introduction

High redshift events are defined as redshifts greater than 4. There are 14 high redshift events and 118 low redshift events. There are a total of 76 features. There are 40 non error features and 36 error features. Many (all?) of the error features come in pairs. For example there is a feature A and then two associated error features A_negerr and A_poserr. Three features (bat_is_rate_trigger, v_mag_isupper, and wh_mag_isupper) take the values yes / no. All other features are continuous.

We fit 6 CART classifiers to the data. There are three costs for misclassifying a high redshift event as a low redshift event: 3, 5, and 7. We apply these three costs to two feature sets: non-error features (40 in total) and all features (76 in total). In section 2 results are discussed. In 3 trees are discussed and displayed.

2 Results

Table 1 displays cross validation error and the redshift value of misclassified high redshifts for each of the six classifiers. The cross validation errors are a bit opaque. With cross validation (specifically 10 fold cross validation), the observations are divided into 10 parts of roughly equal size. One part of the data is "held out" and the model is fit to the other 9 parts. Then the held out part is classified using this model. The error for this held out section is computed. This is repeated for each of the ten sections of the data. "Error" in this case in defined as:

$$\frac{\text{COST * (\# high as low)} + 1 \text{ * (\# low as high)}}{\text{(COST*\# high)} + \text{(\# low)}} \tag{1}$$

I am not sure that this measurement of error is the best way to evaluate performance. Perhaps Tamara has suggestions on something better. Also, I ran the code several times and I am a bit worried about the variability of the cross validation estimate of error. I will try to get error bars on these so the numbers have more meaning.

For each of the 6 classifications, I included the redshift values of the high redshift events that were classified as low redshift under the notion that if most of the error of HIGH as LOW were of events in the 4 - 4.5 range this would be less bad than misclassifications in the redshift ranges of 5 and up. Also seeing which redshift were misclassified gives an idea of whether there is any sort of continuity in the feature space in the sense that one might expect high redshift events just over 4 to be difficult to classify because they are so close to the low redshift events and that as a result most of the misclassifications of high redshift events would occur in the 4-4.5 range. On repeated runs of the code there was a lot of variability in which high redshift were misclassified so do not treat these numbers as revealing any fundamental truths about GRBs.

	Without Error Features	With Error Features
Cost 3	0.731	0.694
	(6.6, 4, 4.3, 4.1, 4.4, 5.1, 8.1)	(6.6, 6.3, 5.3, 4, 4.9, 5.6, 4.3, 4.1, 6.7, 4.4, 8.1, 4.3)
Cost 5	0.665	0.665
	(4, 4.9, 4.3, 4.1, 4.6, 4.4, 8.1)	(6.6, 6.3, 4, 4.9, 5.6, 4.3, 4.1, 4.6, 4.4)
Cost 7	0.681	0.69
	(6.6, 6.3, 4, 4.3, 4.1, 4.4)	(6.6, 5.3, 4, 4.1, 4.6, 4.4)

Table 1: Performance of CART for Two Sets of Features and 3 Loss Functions

3 Trees

Figures 1, 2, 3, 4, 5, and 6 are the trees generated from the data with the three costs and two sets of features used in table 1. These are the trees fit on all the data, not on a cross validated portion of the data. However cross validation is built into the tree growing procedure so in most cases trees do not overfit the data. A new observation dropped into the tree will travel left at a node if the node expression is true, and right if the node expression is false. For example, all trees start with "v_mag_isupper = yes", so if a new observation has yes for "v_mag_isupper" then it will go to the left. The terminal nodes have an "H" if an observation landing in that node is classified as "high" and 'L' if it is classified as low. The two numbers separated by a forward slash at each terminal node are a bit confusing. The left number is the number of high GRB obserbations in that node *times* the cost of misclassifying a high GRB as a low GRB. So in figure 1 the left most terminal node has an 'H' and "33/10". An observation falling into this node will be classified as high. 11 observations from the data in this node where high GRB events, and 10 were low GRB events. In general, if the number on the left of the forward slash is larger than the number on the right, the node will classify as "H", otherwise as "L."

Here are some observations on the trees. v_mag_isupper, A, and flx_pc_late are three important features. Most high gamma ray bursts have 'yes' for v_mag_isupper. The two high redshift gamma ray bursts with no for v_mag_isupper have low FLX_PC_LATE. When v_mag_isupper is yes but A is high (greater than around -1.56) the event tends to be a low redshift GRB. Do these observations from the trees make scientific sense? We could definitely be picking up noise with some of these trees, especially nodes where there are 2 high redshift events. These trees are much more optimistic than the cross validation results in table 1.

3.1 No Error Features

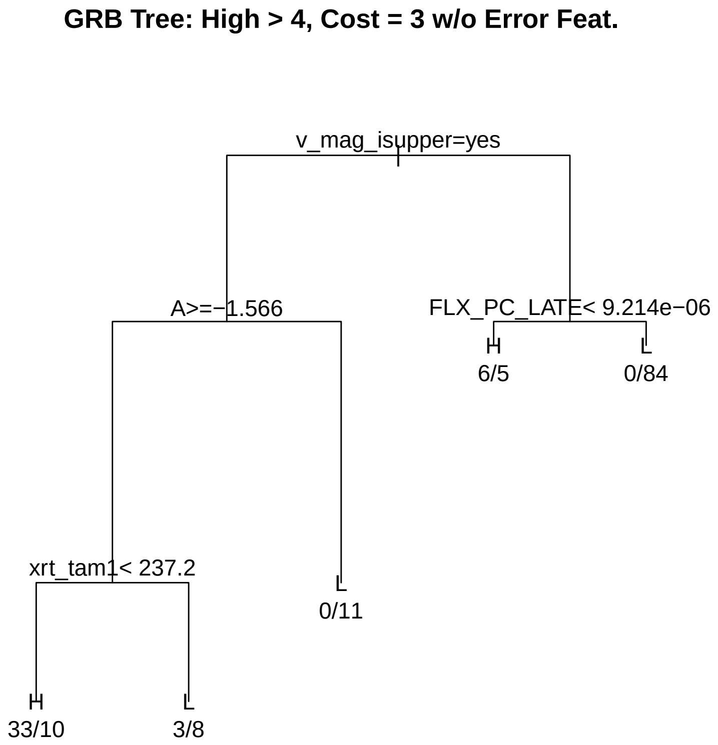
GRB Tree: High > 4, Cost = 3 w/o Error Feat.

Figure 1: Cost 3, no error features.

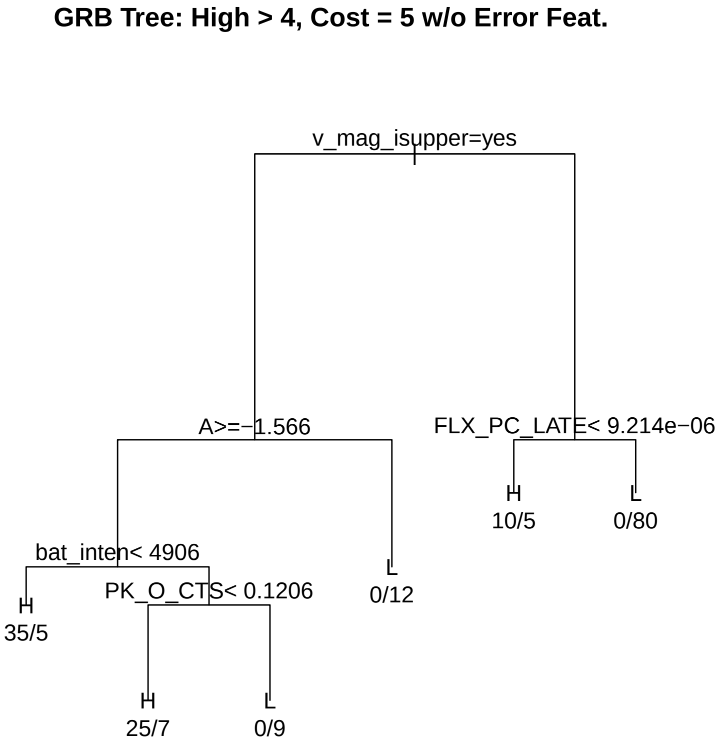

Figure 2: Cost 5, no error features.

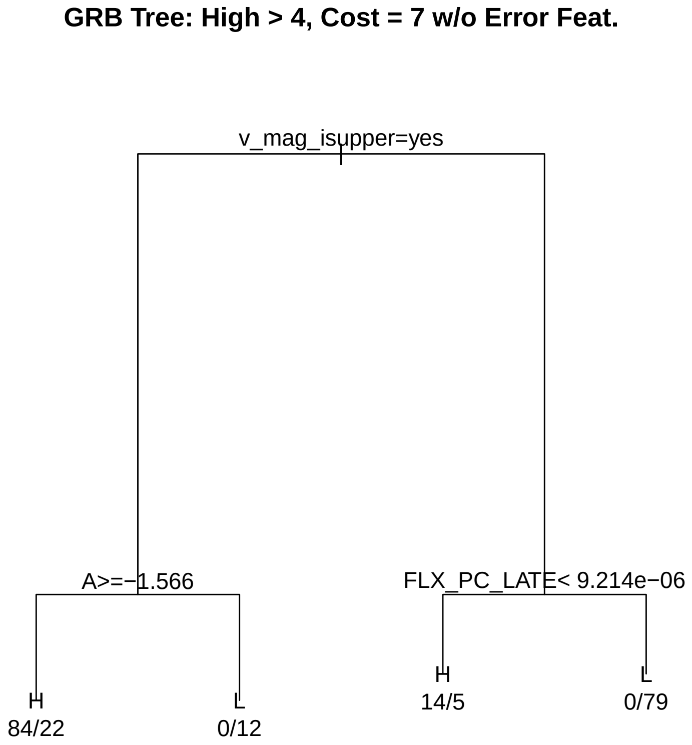

Figure 3: Cost 7, no error features.

3.2 With Error Features

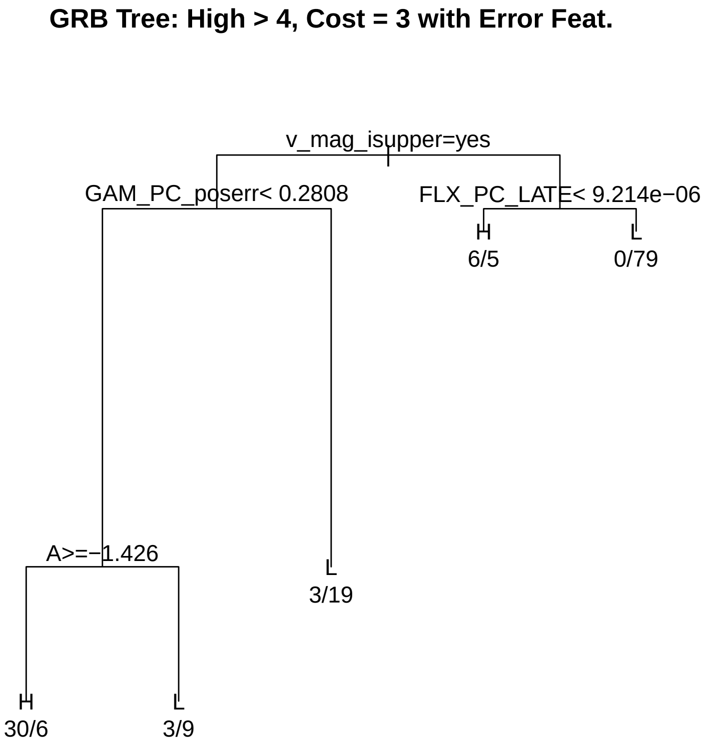

Figure 4: Cost 3, with error features.

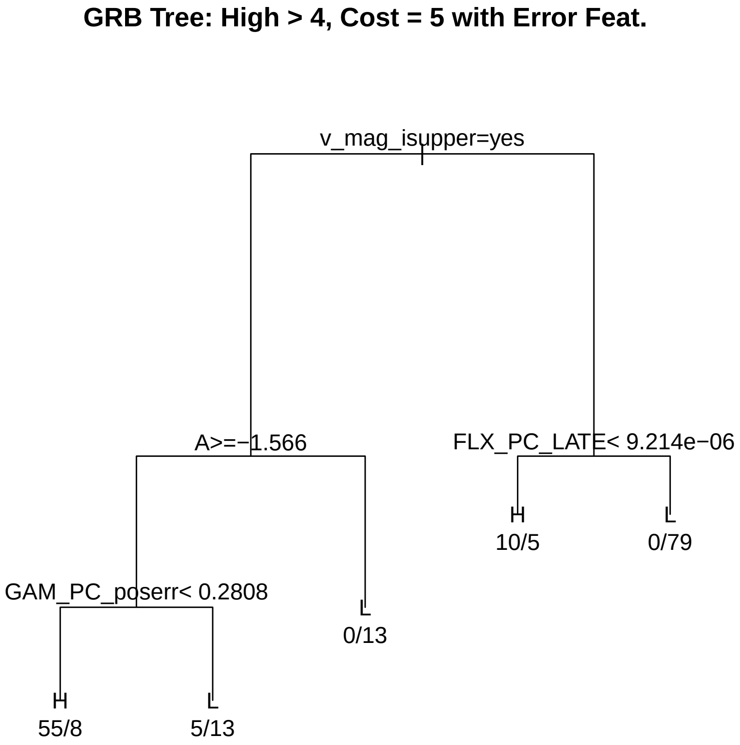

Figure 5: Cost 5, with error features.

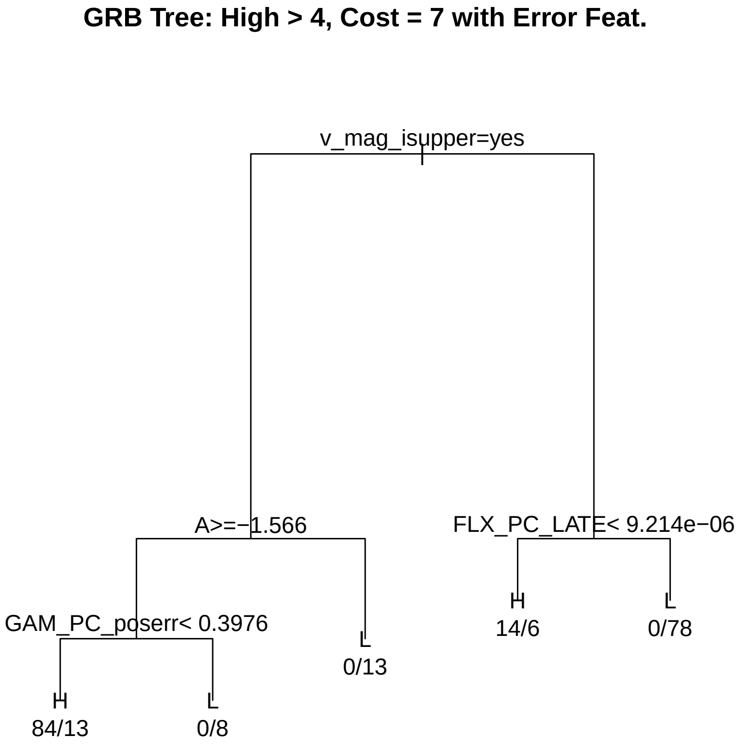

Figure 6: Cost 7, with error features.

4 Some Questions

1. What is the difference between the t90 and T90 features?