

DSC 433/533 – Class 7 – Decision tree example

Task: classify customers as bad or good credit risks (target variable = 1 for bad loan or 2 for good loan) based on:

- banking history = 1 for long and trouble-free or 2 for short and troubled;
- race/ethnicity = Caucasian, African American, Asian American, Hispanic;
- credit card = 1 for bank-issued or 2 for not bank-issued.

Data on 32 customers (15 with a bad loan, 17 with a good loan) available in ClassTree.xls spreadsheet.

Cannot use race/ethnicity information. (Why not?)

Gini coefficient

The gini coefficient for the parent (root) node is $(15/32)^2 + (17/32)^2 = 0.502$.

Consider splitting first on **banking history**, i.e., if history = 1 predict good loan, otherwise if history = 2 predict bad loan.

Since there are 16 customers with history = 1, 2 of which have a bad loan and 14 of which have a good loan, the gini coefficient for the first child node would be $(2/16)^2 + (14/16)^2 = 0.78125$.

Since there are 16 customers with history = 2, 13 of which have a bad loan and 3 of which have a good loan, the gini coefficient for the second child node would be $(13/16)^2 + (3/16)^2 = 0.6953125$.

The weighted average is $(16/32)*0.78125 + (16/32)*0.6953125 = 0.738$, so gini increase would be $0.738 - 0.502 = \mathbf{0.236}$.

Alternatively, consider splitting first on **credit card**, i.e., if credit card = 1 predict good loan, otherwise if credit card = 2 predict bad loan.

Since there are 15 customers with credit card = 1, 6 of which have a bad loan and 9 of which have a good loan, the gini coefficient for the first child node would be $(6/15)^2 + (9/15)^2 = 0.52$.

Since there are 17 customers with credit card = 2, 9 of which have a bad loan and 8 of which have a good loan, the gini coefficient for the second child node would be $(9/17)^2 + (8/17)^2 = 0.50173$.

The weighted average is $(15/32)*0.52 + (17/32)*0.50173 = 0.510$, so the gini increase would be $0.510 - 0.502 = \mathbf{0.008}$.

Since there is a larger increase in the gini coefficient by splitting first on banking history, this is the first split.

Entropy reduction (information gain)

The entropy score for the parent (root) node is $-1[(15/32)*\log_2(15/32) + (17/32)*\log_2(17/32)] = 0.997$.

Consider splitting first on **banking history**, i.e., if history = 1 predict good loan, otherwise predict bad loan.

Since there are 16 customers with history = 1, 2 of which have a bad loan and 14 of which have a good loan, the entropy score for the first child node would be $-1[(2/16)*\log_2(2/16) + (14/16)*\log_2(14/16)] = 0.543564$.

Since there are 16 customers with history = 2, 13 of which have a bad loan and 3 of which have a good loan, the entropy score for the second child node would be $-1[(13/16)*\log_2(13/16) + (3/16)*\log_2(3/16)] = 0.696212$.

The weighted average is $(16/32)*0.543564 + (16/32)*0.696212 = 0.620$, so the entropy loss (information gain) would be $0.997 - 0.620 = \mathbf{0.377}$.

Alternatively, consider splitting first on **credit card**, i.e., if credit card = 1 predict good loan, otherwise predict bad loan.

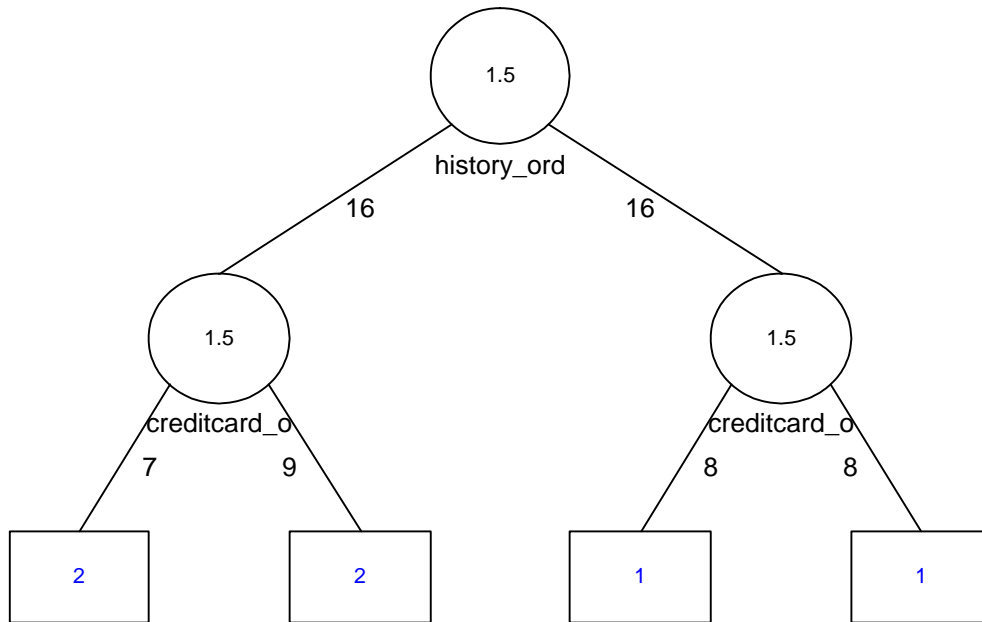
Since there are 15 customers with credit card = 1, 6 of which have a bad loan and 9 of which have a good loan, the entropy score for the first child node would be $-1[(6/15)*\log_2(6/15) + (9/15)*\log_2(9/15)] = 0.970951$.

Since there are 17 customers with credit card = 2, 9 of which have a bad loan and 8 of which have a good loan, the entropy score for the second child node would be $-1[(9/17)*\log_2(9/17) + (8/17)*\log_2(8/17)] = 0.997503$.

The weighted average is $(15/32) * 0.970951 + (17/32) * 0.997503 = 0.985$, so the entropy loss (information gain) would be $0.997 - 0.985 = 0.012$.

Since there is a larger loss in the entropy score by splitting first on banking history, this is the first split.

Full Tree



Rules

- If history = 1 and credit-card = 1 then predict good loan (0 bad, 7 good, so $\text{Pr}(\text{good}) = 1$);
- If history = 1 and credit-card = 2 then predict good loan (2 bad, 7 good, so $\text{Pr}(\text{good}) = 7/9 = 0.778$);
- If history = 2 and credit-card = 1 then predict bad loan (6 bad, 2 good, so $\text{Pr}(\text{good}) = 2/8 = 0.25$);
- If history = 2 and credit-card = 2 then predict bad loan (7 bad, 1 good, so $\text{Pr}(\text{good}) = 1/8 = 0.125$).

Next steps

This decision tree model does well on this (training) sample (2 false positives, 3 false negatives), but would it do as well on new data? To find out, we would need to look at classification errors (and perhaps first decile lift too) in a validation sample.

Another use for a validation sample is to prune the full tree to find one that is not as prone to overfitting. XLMiner can prune the full tree to find a “minimum error” tree with the smallest misclassification error in a validation sample. It can also find a “best pruned tree” which is the smallest tree that has a validation sample misclassification error within one standard error of the minimum error – see class 8 notes and homework 4 for examples.

A final point is that setting the probability cut-off for classification at 0.5 is equivalent to assuming equal costs for both types of error (false positives and false negatives). If there is information about unequal costs for the errors then this could be incorporated into the analysis – see class 8 notes and homework 4 for examples.