CIS 419/519: Homework 1

{Yongxin Guo}

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: Jiangzhu Heng, Yihang Xu

1 Decision Tree Learning

a. Show your work:
Temperature:

$$I(X, Temp) = H(X) - H(X|Temp)$$

And,

$$H(X|Temp) = H(X|Temp \geq 99) + H(X|Temp < 99)$$

With,

$$H(X) = (-\frac{5}{14}log_2\frac{5}{14} - \frac{9}{14}log_2\frac{9}{14}) \approx 0.9403$$

$$H(X|Temp \geq 99) = \frac{2}{7}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) \approx 0.2857$$

$$H(X|Temp < 99) = \frac{5}{7}(-\frac{3}{10}log_2\frac{3}{10} - \frac{7}{10}log_2\frac{7}{10}) \approx 0.6295$$

So,

$$H(X) - H(X|Temp) \approx 0.9403 - (0.2857 + 0.6295) = \mathbf{0.0251}$$

PainLocation:

$$I(X, PainLocation) = H(X) - H(X|PainLocation)$$

As computed before,

$$H(x) \approx 0.9403$$

And,

$$H(X|PainLocation) = H(X|PL = head) + H(X|PL = extremelities) + H(X|PL = chest)$$

With,

$$H(X|PL = head) = \frac{5}{14}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}) \approx 0.3468$$

$$H(X|PL = extremelities) = \frac{5}{14}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}) \approx 0.3468$$

$$H(X|PL = chest) = \frac{2}{7}(0) = 0$$

So,

$$I(X, PainLocation) = H(X) - H(X|PainLocation) = 0.9403 - (0.3468 + 0.3468 + 0) = \mathbf{0.2467}$$

Eventually,

$$InfoGain(PainLocation) = \mathbf{0.2467}$$

$$InfoGain(Temperature) = \mathbf{0.0251}$$

b. Show your work:

Temperature:

$$GainRatio(X, Temp) = \frac{I(X, Temp)}{SplitInfo(X, Temp)}$$

where,

$$SplitInfo(X, Temp) = -\frac{2}{7}log_2\frac{2}{7} - \frac{5}{7}log_2\frac{5}{7} \approx 0.8631$$

So,

$$GainRatio(X, Temp) = \frac{0.0251}{0.8631} \approx \mathbf{0.0291}$$

PainLocation:

$$GainRatio(X, PainLocation) = \frac{I(X, PainLocation)}{SplitInfo(X, PainLocation)}$$

Where,

$$SplitInfo(X, PainLocation) = -\frac{5}{14}log_2\frac{5}{14} - \frac{5}{14}log_2\frac{5}{14} - \frac{2}{7}log_2\frac{2}{7} \approx 1.5774$$

So,

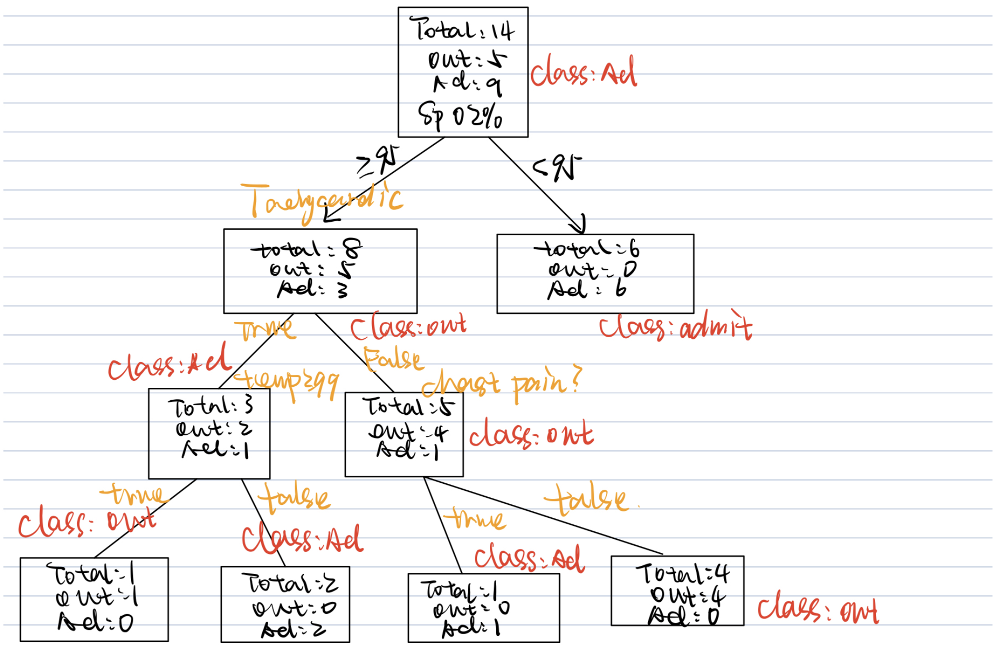
$$GainRatio(X, PainLocation) = \frac{0.2467}{1.5774} \approx \mathbf{0.1564}$$

Eventually,

$$GainRatio(PainLocation) = \mathbf{0.1564}$$

$$GainRatio(Temperature) = \mathbf{0.0291}$$

c.



d. No, ID3 won't guarantee the globally optimal decision tree because it determines the the priority of the nodes based on the information gain. However, looking for the maximized information gain locally will not guarantee a decision tree with globally minimal depth. We can design a set of data that proves this argument by purposely costing the ID3 more paths after it is deceived to choose the root node based on the largest information gain.

2 Decision Trees & Linear Discriminants [CIS 519 ONLY]

A decision tree can include oblique splits by combine all features with one specific relation/equations or so-called constraints instead of evaluating each single feature one at a time. Then a oblique split can be made. For example, we convert all features into numerical values using OHE or some other types of techniques, let's say we have two features X_1 and X_2, then we come out with a split equation, $a_1X_1 + a_2X_2 = C$, where C is a criteria that takes constant value. It should be noted that we must ensure the left hand side of equation is a linear combination of a all the features in order to have a straight oblique split. Then, we can split the data-space by checking if $a_1X_1 + a_2X_2 + ...a_nX_n > or < C$ to implement the partition. Please be noted, the oblique split is a line in 2D, a plane in 3D and a n-dimensional hyperplane in n dimension

3 Programming Exercises

Features: What features did you choose and how did you preprocess them? The features are chosen by applying the correlation filter. The correlation filter is implemented by computing the correlation matrix first and then select the last column where the correlation between diabetic and all the other features are located. The criteria is set to 0.25 as the threshold to filter out the features with lower correlation factor. Eventually the satisfied features are: DIQ050, RIDAGEYR, and BMXWAIST.
Pre-processing was performed as described in the following:
1. The features containing above 25 percent of NaN values are removed
2. Perform the OHE for the nominal features and delete the original nomial features
3. Perform the correlation matrix computation and selected the interested features
4. Replace the NaN value for the rest of columns containing the NaNs 25 percent lower with its the mean of the features.
5. Delete the outliers with 10 s.t.d above or lower than the median

Parameters: What parameters did you use to train your best decision tree
The CCP value is found to be 0.03 as the best/largest value for pruning the trees. The outlier is deleted when the value is 10 s.t.d above or lower than the mean. The NaNs with above 25 percent missing ratio are removed and the rest of columns with NaNs are replaced by the mean of that feature. **Performance Table**:

Feature Set	Accuracy	Conf. Interval [519 ONLY]
DT 1	0.937	0.00106
DT 2	0.935	0.00102
DT 3	0.925	0.00244

Conclusion: What can you conclude from your experience?
The main conclusion we learn from this is that the pre-processing plays a very vital and important role in the accuracy of the model. The feature selection process is also very crucial in terms of generalization of the model. Some of features from the real-world data can be very harmful to the results and performance of the tree. When applying the trained model to the real-world data, one has to make sure that the feature and data structures match with the tree model.