Project 3 Report

CS 4341
Lisa Spalding
Esteban Aranda
9/15/19

Feature Extraction

For this project, 10 total features were utilized. 8 unique features were extracted along with 2 composite features. The two required features included the bottom left corner and which player has more pieces in the center columns. The second required feature was interpreted to mean the center three columns of the board. The six other unique features included: each of the two most centered spots on the board, each of the three center columns, and the bottom row. The first combined feature is both of the center spots, the second is the center column along with one of the two neighboring ones.

The rationale behind choosing the two center spots was to be able to contrast with the lower left corner. If any single spot makes a good feature, it would probably be one of the center two. The combination of the two center features is because there is not a single center spot. The combination of two spaces also makes it more likely to be significant. The reason for the focus on the center in a game of Connect four is because the board has seven columns. Therefore, all winning patterns must have at least one piece in the center row. This also explains the importance of the center columns. To determine how much of a role the center column plays as opposed to the center three, the next few parameters were devised.

The required parameter is which player has more pieces in the center three columns. To determine what within these columns is more important, each column is taken independently as a parameter. This should show whether just the center column is important or if all three are important. The combination of the center column with one of its neighbors should also provide more insight into whether the center column alone is enough or if one of the neighbors is important to provide a better prediction of the result. The row at the bottom is used as a parameter because those would be the first pieces played. If a game is short, the bottom row could end up being very important to the result.

A graphical representation of the features follows on the next page.

Features in graphical form:

			Y			
			Z			
X	X	X	X	X	X	X

Feature 0 is cells Y and Z
Feature 1 is light blue
Feature 2 is center column
Feature 3 is left-center column
Feature 4 is right-center column
Feature 5 is bottom row (denoted by X)
Feature 6 is all three center columns
Feature 7 is center column and left-center or right-center
Feature 8 is cell Z
Feature 9 is cell Y

Feature Evaluation

The best accuracy achieved with a single feature was 81% match with feature 2. Feature 2 was a measure of which player had more pieces in the center column. As stated in the analysis, the center column is the most important in connect 4, so this result is perhaps expected. The column to the right of the center (feature 4) had one of the worst accuracies at 68%, which was rather unexpected. The column to the left of center (Feature 3) had an above average accuracy, at 77%. These facts taken together seem to indicate that while the center column is very important, the columns neighboring it do not have any particular significance. This is supported by feature 6, which is all three center columns taken together, having a slightly over average accuracy at 73%. Feature 7, which is the center column combined with one of its neighbors, has a high accuracy of 78%. Interestingly, this is not higher than the value produced by the center column alone, indicating again that the two columns near the center are not particularly relevant in determining a winner.

The two spots in the center of the center are features 8 and 9. They received accuracy values of 77% and 63% respectively. This shows that no one spot is a good predictor for the winner of the board, especially as the bottom left corner received an accuracy of 78%. It is not expected that the bottom left corner would be significant in determining the winner of Connect four. Feature 5,

the bottom row, received an accuracy rating of 70%. This indicates that the bottom row is not a good way to predict the winner.

The tree was also run using all the features and excluding each feature. Using all the features resulted in a poor accuracy of 67%. The best results for these were when the model excluded feature 3 or feature 8. This is a little counterintuitive, as these features were both fairly accurate on their own. Using more features to evaluate the board resulted in better results overall.

The board was also evaluated with a random forest. The results were different in places, most notably feature 0. Feature 0, the combination of the two center spots, had an around average accuracy in the decision tree, but was very accurate with the random forest method, with a 79% prediction rate. Feature 7, the combination of the center column with a neighbor, was fairly accurate in the decision tree, but was a poor indicator in the random forest at 68%. The bottom row showed the same accuracy of 70%, notably below average for both methods.

The most notable consistency between the two methods was the sustained importance of the center column in determining the winner. With accuracies of 81% and 77% respectively, it was clearly important in influencing the outcome of the game.

Overall, with a board composed of 42 squares, with three possible options for each of those squares, 1000 boards is still a small fraction of possible outcomes. According to wikipedia, there are 4,531,985,219,092 positions in the game. As decision trees are notably bad at extrapolating data over gaps and working with small sets of data, the best accuracy being a relatively low 81% is reasonable in the larger context of the problem.