Will Meyer
CS221

Predicting Football Play Calls

1. Abstract

The ability to predict an opposing team's actions is integral to the game of football. If a team were to know exactly what play their opponents were running in any given situation, they would gain a significant tactical advantage in that they could respond accordingly by calling the proper counter. Currently, teams rely on the collective intuition of their head coaches, defensive coordinators, and on-field players to assess a situation. In my project, however, I built a prediction system that utilizes machine learning algorithms to predict the play type (pass, run, punt, or field goal) that a coach will call in a given situation (down, distance, current score difference, time remaining, field position) on offense. I used the play-by-play data available at pro-football-reference.com as a set of training data, separating data by team, and ran the algorithms for two teams – the San Francisco 49ers and Philadelphia Eagles - in order to determine whether teams showed significant differences in predictability (and whether the approach would work for more than one team). To determine success, I applied the algorithms to 3 games (the most recent ones, from Week 11, 12, and 13 of the 2012 season) that were not included in my training data and determined what percentage of play calls matched the play calls output by each predictive approach.

2. Introduction

As mentioned in the abstract, my goal in this project was to build an algorithm that utilized linear regression techniques to predict an offensive play call in any given situation within a football game. This is essentially the job of every defensive coordinator at any level of football from high school onward; however, their job is somewhat easier as they can see the play formation the other team is utilizing and react accordingly. Additionally, they have personnel knowledge (such as whether a key player has been injured) that can influence their belief as to what play the other team will call. The play predictor, then, essentially does the job of a scout, without the benefit of game film. It can be extremely useful, for example, to know that a given team will pass, say, >80% of the time in a given situation; utilizing regression, we can potentially access this information without having to resort to extensive visual scouting. It is unknown whether this approach has been adopted by any professional teams – if it has, it is likely that their proprietary formulas will be kept secret so as to ensure a competitive advantage, much as MLB and NBA teams do with their sabermetric analysis – but the New York Times recently suggested that NFL teams may be just starting to embrace advanced statistical analysis [6.3], making this a potentially unexplored area of computer science analysis.

3. Approach

To attempt to solve this problem, I used several adapted versions of a stochastic gradient descent algorithm using logistic regression as a loss function, as well as a Naïve Bayes Classifier (with Laplace smoothing). For comparison, I implemented a very simple algorithm: simply predict whichever play type was called most often throughout the training data. The general expectation was that each type of learning algorithm would outperform this baseline.

I did not actually end up implementing a true multinomial regression algorithm for this project, even though the division of data into four classes would seem to suggest such a formulation. My reasoning for doing so was actually quite simple: punts and field goals (two of the four classes) are only called in very specific situations (fourth downs, mostly) and thus can be modeled with heuristics or alternative approaches that allow the problem to be decomposed into a format that allows binary regression classification focused on classifying runs and passes. To that end, I implemented the following algorithms, each of which took in down, distance, yard line (segmented into chunks, based on the assumption that a play run on, say, the 22-yard line is not that much different from a play run on the 20-yard line, but there is a significant difference between, say, the 20 and 50 yard lines), score differential (again, segmented into chunks based on football scoring, which generally comes in chunks of 3 or 7), and time remaining (once again, segmented into chunks):

- Naïve Bayes Classifier: takes in above information, outputs MLE based on training data.
- SGD with heuristics: Uses binary classification SGD to classify runs and passes. Uses heuristics based on down and field position to classify punts and field goals. [7.1]
- One vs. all classification: Four separate binary classifiers were trained using one play type as positive and all others as negative; prediction was done using the classifier with the highest prediction score. This allowed classification to account for punts/field goals without the need for heuristics.
- All vs. all classification: Six separate binary classifiers were trained using only data from two types of plays (i.e. run vs. punt, run vs. field goal, run vs. pass, etc.) each, then each classifier "voted" on the most likely play; the classifier output the play type with the most votes. This allowed a separate, more complex feature extractor to be used for the run vs. pass classifier; all others achieved >98% success rate using only basic input data.

4. Experiments/Analysis

For each team, I used roughly a season and a half of training data (taken from the 2011 and 2012 seasons) and used the three most recent games from the 2012 season as validation examples. Data complexity/time was not an issue as the number of training examples was limited by necessity (go back too far in a team's history and personnel turnover – players, coaches, etc – makes predictions based on past behavior meaningless; furthermore, data based on more than one team would be meaningless given the difference in personnel). The results, described in the appendix [7.2], showed a marked improvement over the baseline classification – in particular, with the most accurate method (all vs. all classification), the 49ers data displayed over 72% accuracy on validation data, compared to the baseline of 41.4%. Perhaps the most encouraging aspect of the algorithms was their ability to classify certain situations extremely well. The SGD-based models excelled in classifying punts, field goals, and down/distance situations outside of 1st and 10, displaying 82.27% accuracy in these situations for Philadelphia and 88.84% accuracy for San Francisco. Naïve Bayes classification revealed the opposite; its lack of situational awareness in nuanced situations accounted for most of the accuracy difference between its output and that of the SGD models (for example, on plays with a longer distance such as a 2nd and 10, it would predict a pass even if the team was ahead and trying to run out the clock because the relative probability of a longer distance given pass dwarfed the relative probability of a small, but positive score differential given run).

Examining the misclassifications made by the SGD algorithms for both teams reveal several trends that identify the remaining weaknesses in the model (for simplification, we examine only mistakes made by the run/pass classifier; accuracy on field goals and punts was in the high 90s):

- 1st and 10: The plurality of the mistakes made by the SGD classifier were made on plays run on 1st and 10, the starting position for every drive in football. Of the 141 total mistakes made on validation data, 62 were made on plays occurring on 1st and 10, accounting for almost 44% of the total error. Some of this can be accounted for by the fact that 1st and 10 is the most common down/distance situation in football. In fact, between the two teams, there were 156 total occurrences of 1st and 10 in validation data. Only 60.26% of these were classified correctly, though. A decrease in overall accuracy, however, is somewhat expected in this circumstance; there are relatively few occurrences of this situation in which the play call is anything more nuanced than an educated coin flip (i.e. the situation "1st and 10", in general, dictates neither run nor pass).
- Quarterback runs: Plays where the quarterback runs for positive yardage are recorded in box scores as runs, despite the fact that the play call on these types of plays is usually actually a pass. Unfortunately, these errors are not correctable without access to game film, as quarterback sneaks/draws/zone reads are legitimate running play calls that are recorded the same way in a box score. These accounted for 10/141 (>7%) of validation error, but may have been responsible for more, as these errors also occur in training data. [3][4]
- Distances longer than 10 yards: These accounted for 13.48% of validation error. Classification may be inaccurate in these cases because they are relatively rare; they require either a loss of yardage on the initial play of a series or a penalty, restricting the number of training instances used for regression. [5]

Most other misclassifications range from simple misfires [6] to edge cases [7] to the bizarre [8]; these phenomena seem to be the most apparent ones.

5. Conclusion

While regression analysis displayed a certain degree of predictive accuracy, it is only one part of a small puzzle. It is not a foolproof way to predict a coach's behavior, and absent visual evidence, such as the formation a team in which a team lines up on a given play, it is, in fact, an inherently flawed one; additionally, some part of the analysis predicts known outcomes (for example, every football fan knows a team down by one score with less than two minutes left to play is probably going to be passing on every play). However, there is apparent value in analysis of subsets of the data, specifically 2nd and 3rd down and distances less than 10 yards (as predictions in these instances appeared to be particularly accurate). For example, knowing that a particular team has an 80% chance of passing on, say, 2nd and 5 presents a definite strategic advantage to the defense; it would also be helpful to know that a team only runs, say, 75% of the time in a situation where the rest of the league runs 90% of the time, as it would allow the defense to prepare for a relatively unexpected outcome. An obvious next step would be to run analysis on more teams and see if some can be isolated as particularly predictable; other additions to the process include the potential inclusion of more factors (such as the formation a team is using) and finding a way to correct for mislabeled plays (quarterback scrambles or fumbles classified as runs).

6. References

1. Philadelphia box scores: http://www.pro-football-reference.com/teams/phi/2012_games.htm, http://www.pro-football-reference.com/teams/phi/2011_games.htm
2. San Francisco box scores: http://www.pro-football-reference.com/teams/sfo/2012_games.htm, http://www.pro-football-reference.com/teams/sfo/2011_games.htm
3. Sabermetrics in football: http://www.nytimes.com/2012/11/25/sports/football/more-nfl-teams-hire-statisticians-but-their-use-remains-mostly-guarded.html?_r=0

7. Appendix

1. Note: combinations of input data were used as features to increase accuracy of this model. For example, the combination of down/distance/score differential as a single feature proved to be a strong predictor of run/pass behavior relative to using the individual factors as features. Intuitively, this makes sense, as coaches don't consider each variable in a vacuum when calling a play. The fact that this type of combination approach proved accurate also speaks to why the Naïve Bayes classifier was less accurate, as considering each variable independently was exactly the methodology behind its implementation.

2.

Algorithm	Accuracy (Philadelphia Eagles)	Accuracy (San Francisco 49ers)	Accuracy (average)
Baseline (most common play)	55.67%	41.40%	48.53%
Naïve Bayes Classifier	59.61%	52.09%	55.85%
Naïve Bayes Classifier w/Laplace (1)	58.62%	53.49%	56.05%
Stochastic Gradient Descent w/heuristics	62.07%	70.23%	66.15%
One vs. All binary SGD	66.01%	64.65%	65.33%
All vs. All binary SGD	62.56%	72.09%	67.33%

3. Concrete example: 2,13:03,3,10,RAM 42,Colin Kaepernick for 1 yard (tackle by James Laurinaitis),7,0,1.55,0.53,False,SFO (data in format: quarter, time, down, distance, yard line, play, away score, home score, (ignore), (ignore), is team at home?, team)

It is highly unlikely that Kaepernick was actually supposed to run on this play given that the 49ers were right on the edge of field goal range, needing 5 yards for a reasonable field goal attempt and 10 yards for a first down. It is far more likely that he dropped back to pass and later scrambled, yet the box score reflects it as a run (and the predictor as a missed prediction).

4. It is worth noting that the Eagles quarterback for the last year and a half, Michael Vick, is known for scrambling often when plays break down. This might explain a good deal of the lowered accuracy in the Philadelphia play predictor.

5. 1,8:08,2,17,RAM 44,Colin Kaepernick pass complete short left to Bruce Miller for 11 yards (tackle by Jo-Lonn Dunbar),0,0,1.63,2.41

Without scanning all 2000+ training examples for San Francisco, I doubt there are more than one or two 2nd and 17s. I tried to alleviate this problem by grouping distances from 11-15 yards, 16-20 yards, and 20+ yards.

6. Concrete example: 1,4:18,2,4,RAM 8,Frank Gore right guard for no gain (tackle by Quintin Mikell),0,0,5.09,4.28,False,SFO

2nd and 4 indicates they should run, maybe; field position dictates they should pass to try and get a touchdown, or even the first down (again, maybe); you could make an argument for either. The predictor margin was quite thin – it just predicted wrong.

7. Concrete example: 4,8:42,4,1,CAR 40,Bryce Brown right end for no gain (tackle by Frank Alexander),24,22,0.59,-1.6,True,PHI

This is a situation where the Eagles are just inside field goal range, but the probability of making one isn't very good and they don't gain that much of a field position advantage by punting. You can make arguments for calling a pass, run, field goal, or punt here. For what it's worth, the All vs. All approach suggests that they are most likely to run (the actual outcome of the play) if they don't punt, dictating a proper course of action once they lined up in a non-punting formation.

8. Concrete example: 3,1:04,4,7,RAM 14,David Akers pass complete short left to Michael Crabtree for 14 yards touchdown,26,10,2.51,7,False,SFO

In case you're not familiar with the 49ers roster, that's the kicker throwing a touchdown pass, presumably on a fake field goal. (this is actually a training error, but it's also an excellent example of a strange play)