presented at the MIT Sloan Sports

Analytics Conference on March 6, 2010

Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing

Joseph Sill
joe@hoopnumbers.com

Abstract

Adjusted +/- (APM) has grown in popularity as an NBA player evaluation technique in recent years. This paper presents a framework for evaluating APM models and also describes an enhancement to APM which nearly doubles its accuracy. APM models are evaluated in terms of their ability to predict the outcome of future games not included in the model's training data. This evaluation framework provides a principled way to make choices about implementation details. The enhancement is a Bayesian technique called regularization (a.k.a. ridge regression) in which the data is combined with *a priori* beliefs regarding reasonable ranges for the parameters in order to produce more accurate models.

Adjusted +/- (APM) is a player evaluation technique which is growing in popularity within the basketball statistics community. APM, which was invented by Wayne Winston [5] and Jeff Sagarin and first described in detail by Dan Rosenbaum [4], starts with the simple observation that what matters is the player's contribution to the team's margin of victory. The first thing

which comes to mind is to track the +/- stat, which is the difference between team points scored and opponent's points scored while the player was on the floor. The problem with using +/- to evaluate a player is that it doesn't take into account who the player played with and played against.

APM is a method which uses regression to take into account the teammates and opponents while a player was on the floor. A regression is run on a large dataset of games, where each data point in the dataset is a game snippet during which there were no player substitutions. The regression produces a rating for each player, roughly representing the effect the player is predicted to have on a team's margin of victory if that player were to replace an average NBA player in the team's lineup.

How well does adjusted +/- actually work? This has been a vexing and controversial question within the basketball analytics community. Certainly, the results are not always intuitive, particularly when the analysis is run on a single season's worth of data.

Why might APM be giving such non-intuitive results in some cases? Is it really discovering that the true value of various players is dramatically different from their commonly perceived value? In some cases, it may be doing so, but there are reasons to be concerned about APM's accuracy. Within the basketball analytics community, much attention has been focused on the issue of multicollinearity, which in this context corresponds to situations where pairs of players are very frequently or very rarely on the floor at the same time. However, it is the opinion of the author that multicollinearity is only part of the story behind the struggles of the APM technique. There is a more general phenomenon at play, known amongst the statistics and machine learning community as *overfitting* [1].

Overfitting occurs when a model fits the training data too precisely, i.e., in such a way that the fluky pecularities and noise of the data are fit, so that the model's predictive performance on future data is degraded. As a simple example, imagine a situation where a rookie NBA player has a FG% of 67% over the first 5 games of the season. Suppose the task is to predict the player's FG% over the course of the remaining 77 games. It should be intuitively obvious that the estimate which best fits the available data, 67%, is unlikely to be the optimal prediction for the entire season. The naive use

of such an estimate for the purposes of prediction would be an instance of overfitting. A smarter approach would be to combine the data from the 5 games with prior information about the typical distribution of FG% which NBA players achieve over the course of the season.While it is true that APM models are often estimated over an entire season of data (or even multiple seasons) it is nonetheless the case that standard linear regression is prone to overfitting, since the data is so noisy and since so many parameters (one for nearly every player in the league) need to be simultaneously estimated.

Another challenge regarding APM surrounds various implementation details where choices have to be made, apparently somewhat arbitrarily. Some practitioners lump all players whose total minutes played is below a threshold into a single "reference player", while others advise against this. There is also disagreement about whether to use multiple seasons of data and how to weight it when used.

This paper proposes a technique for evaluating APM and choosing the proper settings for implementation details (it also presents a method for combatting overfitting, which will be discussed later). The evaluation technique is both widely used by predictive modeling practitioners and intuitively reasonable given the way that APM is used. An NBA front office might run an APM regression on historical data and then make a decision influenced by the APM results regarding a trade or free agent signing. Then, the team will have to live with the results once further games are played with the new acquisition. Therefore, what we'd like to know is APM's ability to predict what will happen in games which are *not* included in the dataset the APM model was fit on.

Fortunately, there is a straightforward procedure often used in applied statistics for assessing a model's ability to make predictions. The model is fit on most of the data, with some of the data reserved for testing only. More generally, one can split the data many different times, each time fitting on most of the data and predicting on the rest, and average the test results over many different splits. This is a well known technique in statistics and machine learning known as cross-validation (CV).

The specific testing framework used in this paper is as follows. Given the game's substitution history (i.e., the data regarding who was on the floor vs.

whom and for how long), the APM model is used to generate a prediction about the outcome of each game snippet in the game. Specifically, the prediction is the difference in efficiency (points per 100 possessions) of the home team versus the road team during that snippet. The prediction is then scaled by the number of possessions which occurred during the game snippet to get a prediction of the scoring margin occurring during the game snippet. Finally, the game snippet predictions are summed to get a prediction about the overall margin of victory during the game. The predicted margin of victory is then compared to the actual margin of victory. The accuracy of the model on a collection of test games is measured by the root-mean-squared-error (RMSE) of predicted vs. actual margin.

CV can be used not just to evaluate a model, but also to determine the best settings of "meta-parameters" such as the minutes cutoff (for the reference player) and the emphasis placed on past years. The meta-parameter settings which yield the lowest CV RMSE are chosen, since those settings result in the best prediction accuracy. In order to get a rigorous evaluation of the model, however, some data was set aside for pure testing at the very end. This data was not used in the cross-validation experiments, so the RMSE on this data is an unbiased estimate of the true RMSE given the meta-parameters chosen by CV.

CV was used to determine the optimal minutes cutoff, M, and weightings of past years for the standard APM linear regression technique. 3 years of NBA play-by-play data was available, ranging from the '06-'07 season to the '08-'09 season.The data from the 2008-2009 season was split it into 2 sets: the games through February, and the games from March and April. The March and April games were used as the pure test set, so they were not included during the cross-validation phase for tuning the meta-parameters. Note that the trade deadline is usually in late February, so this split roughly corresponds to the amount of data a team would have collected just before the trade deadline. The games through February are split into 10 equisized subsets in order to do cross-validation. The average of the results on the 10 subsets is the CV RMSE.

In order to try to assess the value of using prior data and the right way to weight it, results when the model is fit only on 2008-2009 data were

compared to the results when 2007-2008 and 2006-2007 data is also used. The previous two years' data was always in the training set for every CV split. We use a weighting scheme in which data from k years ago is weighted with weighting param D^k, for various values of D.

The standard deviation of the margin of victory amongst the March/April test set games is 12.58, so an RMSE of 12.58 represents an R-squared of 0, i.e.,a trivial accuracy level which could be achieved by predicting the mean home margin of victory of 3.61 points.

Using only the '08-'09 training data through February, CV RMSE was minimized for M=1200 minutes when using standard linear regression-based APM. Unfortunately, the test RMSE of this model with M=1200 on the March/April test set is 12.76, or worse than is achieved by the trivial home-court-advantage model. Such poor out-of-sample prediction performance clearly indicates a great deal of overfitting.

Using three seasons' worth of data while using standard linear regression, the optimal meta-parameter settings were M=800, D=0.75. The test RMSE of this model was 12.01 (R-squared 0.088). Thus, extra seasons of data prove highly beneficial when using standard linear regression, but the test R-squared still remains somewhat disappointing.

In order to combat overfitting and improve accuracy, a technique known as ridge regression [3, 2] (also known as Tikhonov regularization) was applied. Regularization is a widely-used technique in the statistics and machine learning community and has applications far beyond linear regression. In this context, using regularization amounts to minimizing

$$\sum_i p_i(t_i - \mathbf{w} \cdot \mathbf{x})^{\mathbf{2}} + \lambda ||\mathbf{w}||^{\mathbf{2}} \tag{1}$$

where p_i is the number of possessions in the game snippet, t_i is the difference in offensive efficiency between the home and road teams for the game snippet, $\mathbf{x}$ is a vector of 0s,1s, and -1s with the 1s representing the presence of the home players on the court and -1s representing the road players (there is also a constant 1 representing the ever-present home court advantage). $\mathbf{w}$ is the vector of player ratings which are to be estimated. Standard linear regression corresponds to minimizing the first sum only without the second term penalizing large parameter values.

There is a useful Bayesian interpretation of this technique. The second term can arise out of a Bayesian *prior distribution* over $\mathbf{w}$ of independent gaussians. The λ in the equation is corresponds to the ratio of the variance of the inherent, unpredictable noise to the variance of this gaussian prior.

The optimal λ was chosen via cross-validation as with M and D before. For three years of data, the optimal combination of meta-parameters was λ=3000, M=1200, and $D = 0.25$. This model achieves a test RMSE of 11.47 (R-squared 0.169).

When using regularization, the benefit of additional years of data and the use of a minutes cutoff turns out to be modest. With just 1 year of data and M=0, the optimal λ is found to be 2000 and the corresponding test RMSE achieved is 11.54 (test R-squared 0.159), which is nearly as good as the best result achieved with 3 years of data.

It is possible to deduce that a λ of 3000 corresponds to a gaussian prior with standard deviation 2.71, while λ=2000 corresponds to a gaussian prior of 3.32. A gaussian prior with standard deviation around 3 implies that 95% of the players *a priori* are presumed to be within the range $[-6, 6]$, which seems intuitively reasonable, bearing in mind that 1 point per game in margin of victory over the course of an 82 game season is worth nearly 3 wins on average.

The improved accuracy when using regularization is not merely a fine-tuning of the results using standard linear regression. Dramatic changes in the rankings of players can occur. For instance, in a comparison of 1-year '08-'09 results with and without regularization, Steve Blake's ranking is boosted by regularization from the 22nd percentile to the 73rd percentile, and Delonte West moves from the 38th percentile to the 87th percentile.

Space considerations prevent further elaboration, but there are additional results available at hoopnumbers.com (the author's website) which may be of interest. For instance, there are regularized adjusted plus/minus results which isolate a player's impact on team offensive and defensive performance regarding the four factors (effective FG%, free throws, rebounding and turnovers) defined by Dean Oliver.

It is the author's belief that the themes presented in this paper- the focus on out-of-sample predictive accuracy and the use of Bayesian techniques-

are broadly applicable to many problems tackled by the sports analytics community.

References

[1] Overfitting. http://en.wikipedia.org/wiki/Overfitting.

[2] Tikhonov Regularization. http://en.wikipedia.org/wiki/Tikhonov_regularization.

[3] A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.

[4] Dan T. Rosenbaum. Measuring How NBA Players Help Their Teams Win, April 2004. http://www.82games.com/comm30.htm.

[5] Wayne Winston. Player and Lineup Analysis in the NBA, New England Symposium on Statistics in Sports 2009, September 2009. http://video.yahoo.com/watch/6196780/16087711.