Discrimination and Clustering – Can we learn from this college football data set?

Scott Snyder, Neepa Subramanian and Jiayang Sun
Case Western Reserve University

Abstract

Football is an important part of American life. One important aspect of any team is the physical characteristics and performance abilities of its starting players. In this research we explore the relationship between some qualities of players and their starter/non-starter membership. In other words, we are interested in finding a good discriminant rule for classifying individuals' membership based upon some (primarily physical) dependent variables, for each of the different field positions.

A challenging issue is that what is deemed "important physical characteristics", and the methods by which these characteristics are measured, vary by team, school and maybe even conference. The result is that the data come with many missing values on different variables in different ways depending on teams and conferences. Efficient and realistic imputation algorithms are sought. Here the use of a regression approach as well as a neural net is presented.

Pretending that we do not know the membership, a cluster analysis can be performed to see if there is a reasonable grouping of the players into clusters of starters and non-starters for each of the playing positions. This is helpful to determine the quality of a discrimination rule based on the available data and may suggest extra factors to be considered in training athletes.

When the data dimension is high, there is a curse of dimensionality in searching for groups. Thus, we use projection pursuit, clustering techniques and CART in conjunction with knowledge about the requirements for each position to select a set of variables which is most likely to show important structures. It is very interesting to see if the line found by the Fisher's linear discriminant rule is orthogonal to the "interesting" direction suggested by projection pursuit.

AMS 1991 subject classifications. Primary 62H40, 62J99; secondary 62H30, 82C32.

Key words and phrases. Missing values, imputation, CART, neural nets, logistic discrimination, projection pursuit and cluster analysis.

Short Title: Discrimination and Clustering

Address and Phone No: Department of Statistics, Case Western Reserve University, Cleveland, OH 44106. 216-368-0630 (Tel), 216-368-0252 (Fax), jiayang@sun.cwru.edu.

1 Introduction

The data collected involves physical characteristics and performance measures on over 1700 college football players who play at one of 17 positions. In addition, we are also given whether or not the player was a starter for the team. See Table 1. Of primary interest in this analysis is to 1) examine the relationships between these variables; 2) utilize all the data by effective and reasonable treatments of missing values; 3) compare discrimination techniques and dimension reduction techniques (including clustering techniques) (cf. [4], Everitt and Dunn, 1992); and 4) see if Fisher's linear discriminant rule is orthogonal to an "interesting" direction suggested by projection pursuit (cf, [12], Sun, 1998). Interest 3) is an attempt to predict starter/non-starter status with reasonable probability, based on each player's physical and performance measures. The cluster analysis is helpful to determine the quality of a discrimination rule based on the available data and may suggest extra factors to be considered in training athletes. The discrimination rule will be developed separately for each position as it was reported in [1] (Berg et al, 1992) that players in different positions have differ-

ent physical attributes.

Since coaches are always interested in what makes the best players, we hope to provide some insight into some physical characteristics that starters had at the time of data collection. This information could then be utilized by non-starters to identify specific areas of weakness.

From a statistical point of view, this complex dataset offers the opportunity to examine a highly missing situation and utilize projection pursuit to find potentially interesting structure. Here we would like to find a bimodal projection that leads to a reasonably good separation between starters and non-starters and compare that projection to Fisher's LD rule.

Thirteen variables that represent a wide range of physical characteristics were collected. Those variables are height, weight, percent body fat, vertical jump, long jump, 40 yard dash, 20 yard dash, and one repetition maximum on the bench press, incline bench press, overhead press, power clean, back squat, and leg press. These variables are consistent with some previously reported studies; see [1] (Berg et al, 1992), [3] (Craft, 1992), [5] (dos Remedios, 1992), and [6] (Bridgman, 1991) for some examples. Although all the variables were requested, only the information that was readily available was provided by each school. In addition, no attempt was made to control for weight lifting technique. Classification variables to distinguish between conferences and positions will also be utilized.

Table 1: Variables in the Football Data Set

Status	Positions (17 positions)
Conference	Height (inches)
Weight (pounds)	Percent fat
Bench press (pounds)	Inclined bench press (pounds)
Overhead stress (pounds)	Power clean (pounds)
Back squat (pounds)	Leg press (pounds)
Vertical jump (inches)	Standing long jumps (inches)
40 yard dash (seconds)	20 yard dash (seconds)

2 Exploratory Data Analysis and Imputation

Initial exploration of the data shows position 17 (place kicker) has only four observations. In addition, approximately 185 observations have the position variable missing. Since our subsequent analysis is dependent upon the variable position, these observations will be dropped. The resulting dataset has 16 positions and 1518 observations.

The frequency of missing data in this dataset is quite extensive, ranging from 11% missing in the 40 yard dash to 98 % missing in the incline bench press. This is immediately disturbing for this cursory examination would lead us to expect to get little information from incline bench press and other highly missing variables. Thus, an effective and sound treatment of missing values is essential. Our treatment of missing values follows the following principles (*after investigating why they are missing and if there is any pattern of missing*):

- If they can be treated as *missing at random* (MAR), delete *cases* if the percentage of missing is small; delete the *variable* if the percentage of missing for this variable is close to 100% (for analyses, though may not for imputation); impute missing values using a *good* imputation method (e.g. iterative, multiple and multivariate imputation techniques) in the intermediate cases. Use a combination of deletion and imputation and other knowledge in more complicated cases. Try analyses with missing values imputed *and* deleted, respectively, if possible.

- If there is a pattern of missing that depends on the true values of the variable(s) of interest, we have a *biased sample*. The biased sample is from a population different from the target population. An analysis or imputation ignoring the bias thus often leads to ridiculous conclusions. Either a simple treatment (cf. [8], Patil and Rao, 1977, and [13], Sun, 1998) or a more sophisticated treatment (cf. [15], Vardi, 1982, [14], Sun and Woodroofe, 1997, and [7], Manski, 1993) for biased samples must be applied in this situation.

- And, of course, we should collect more data to make it more complete if possible.

Xgobi (cf [11], Swayne, Cook, and Buja, 1998) and Splus graphics are useful for further examination of the pattern of missing values. The missing values in this data set are mostly related to conferences and loosing speaking, they can be treated as MAR. Both a regression and neural network approach was taken in order to estimate or impute missing values. The regression approach used was Minimum Generalized Variance (MGV) from SAS's PROC PRINQUAL [10]

and the neural net utilized was by Venables and Ripley (1997) in Splus. See [16] or [9].

The first decision was to not even attempt to use the variables incline bench press (98% missing) percent body fat (87% missing), overhead press (91% missing), leg press (88% missing), long jump (86% missing), and 20 yard dash (82% missing), because they were so sparsely present. The resulting imputation was very unsatisfactory, and both the MGV method and the neural net gave unreasonable predictions (eg. heights$<$ 3 ft., squats $<$ 0, etc.). When all of the thirteen variables were included, the MGV method gave reasonable results. The authors have not yet attempted to write a modified neural net imputation algorithm to effectively impute the missing values using all the data in this complex dataset, in the spirit of multiple iterative non-linear imputation.

Even though the MGV method gave reasonable results, the imputed values in the highly missing variables are still not useful. As one would expect, most were given approximately the same value. However, this data was extremely useful in predicting values of the more complete variables. The reasoning is this; for example, in order to predict missing values of squat we'll use all the other available information for the individual to do so. In an optimistic situation, the variable leg press would be available. Leg press should be a much better predictor of squat than, say, height. See Figure 1.

As a final introductory note, weight was transformed to log(weight) to correct for skewness. This new variable will be referred to as weight throughout the remainder of this analysis.

3 Analysis

3.1 CART and Discrimination

CART (Classification and Regression Trees) by [2] (Breiman et al, 1984) was used as the first discrimination tool. As it turns out, CART (based on cross validation) had a slightly better error rate than that of logistic discriminant analysis, and did considerably better than Fisher's linear discriminant rule, and quadratic discriminant rule. The final list of variables used were height, weight, bench press, squat, vertical jump, and 40 yard dash, although for each position, only some of these variables were important and that too in different orders. Table 2 provides a summary of the important variables by position for the CART discrimination.

Figure 1: Two Scatter Plots

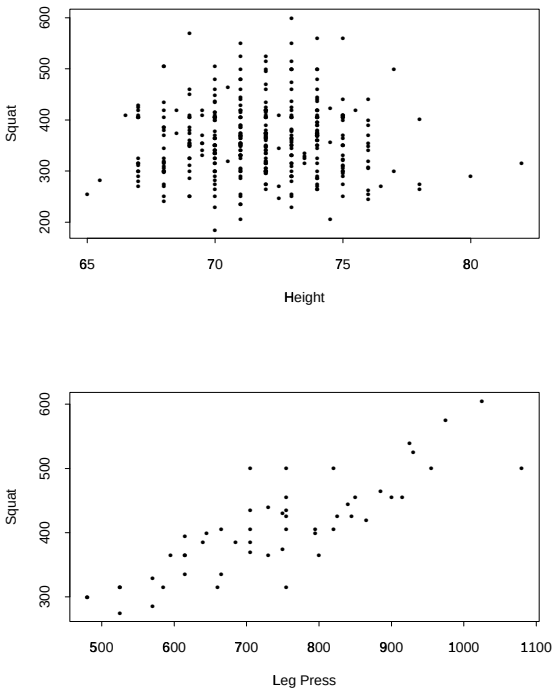


Table 2: Summary of Important Variables

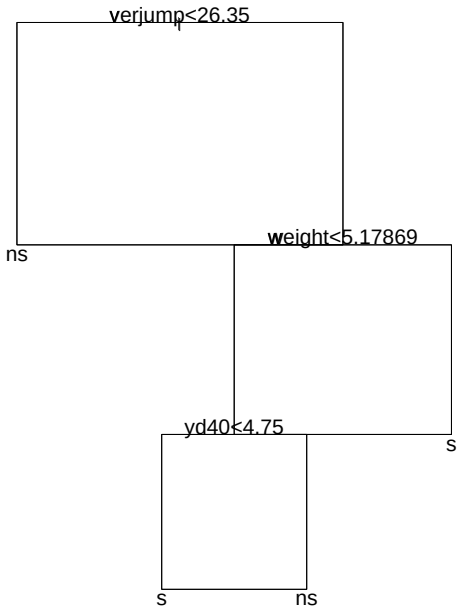
Position	Variables
1. Nose Tackle	squat, 40 yd dash
2. Defensive Tackle	vert. jump, squat
3. Defensive end	bench, weight
4. Outside Linebacker	bench
5. Inside Line backer	squat, weight
6. Cornerback	bench, 40 yd dash
7. Free Safety	vert. jump, weight, 40 yd dash
8. Strong Safety	40 yd dash, weight, squat
9. Offensive Center	squat
10. Offensive Guard	bench, squat, 40 yd dash, weight
11 Offensive Tackle	bench
12. Tight End	bench, squat, height
13. Wide Receiver	bench
14. Quarterback	bench, weight, height
15. Fullback	squat, 40 yd dash ,weight
16. Running Back	weight, vert. jump, bench

The size of each tree was determined by the change in deviance plot. In most cases the deviance began to go up beyond a size of 2 or 3. This indicates that for most positions one or two variables are all that is necessary to distinguish between starters and non-starters. Although some positions suggested only a size of 2, a minimum of 3 was used. This is in order to make some of the trees appear more reasonable.

Table 3 compares the probability of correct discrimination between CART and Logistic discrimination. The overall probability for all 16 positions is .769 for CART and .756 for Logistic discrimination.

Figure 2 shows an example of the classification tree for the position 7 (Free Safety). For CART, this position had the best classification rate at 86%. The tree is also reasonable for what is known about the duties of a free safety. The tree suggests that the starters in this data can jump higher, weigh more, and are generally faster then their non-starter backups. This seems to make sense in that the player performing the job of a free safety is often in a position to perhaps intercept or deflect a pass, tackle an opposing player, and certainly a must also be able to run a quick 40 yard dash.

Figure 2: Tree Diagram for Free Safety



All 16 trees and 16 logistic discriminant rules will be made available on a web page for public access after the presentation in Joint Statistical Meetings at Dallas in August, 1998.

3.2 Projection Pursuit

As previously mentioned, Fisher's linear discriminant rule did not give as small of an error rate as CART or logistic discrimination. This is expected since most classification variables are not normally distributed. It is interesting, however, to see whether or not the solution by Fisher's linear discriminant rule is (almost) orthogonal to an "interesting" direction found using Projection Pursuit (PP). PP looks for interesting projections that can reveal most interesting structures of the data, such as clusters, bumps and other nonlinear structures. A normal distribution is often viewed least interesting in PP (cf, e.g. [12], Sun, 1998). Of primary interest here would be to find a projection that results in a bimodal distribution separating the starters and non-starters. This was explored by position and, visually speaking, reasonable projections could be found for most of the positions. A histogram of an interesting projection for the position Free Safety is shown in Figure 3. For some projections that looked interesting, the efficiency of classification by that particular set of projection coefficients was not as good as CART or logistic discrimination. This is also expected since the PP tool used was some sort of dynamic projection pursuit that allows a user interface and looks for more than just clusters.

Table 3: Classification Performance

Position	#	CART	Logistic
1. Nose Tackle (NT)	32	.781	.813
2. Defensive Tackle (DT)	124	.742	.750
3. Defensive end (DE)	107	.776	.813
4. Outside Linebacker (OLB)	103	.748	.728
5. Inside Line backer (ILB)	103	.777	.825
6. Cornerback (CB)	122	.795	.762
7. Free Safety (FS)	50	.860	.720
8. Strong Safety (SS)	58	.810	.741
9. Offensive Center (OC)	63	.762	.714
10. Offensive Guard (OG)	126	.770	.722
11 Offensive Tackle (OT)	118	.695	.720
12. Tight End (TE)	73	.795	.767
13. Wide Receiver (WR)	166	.741	.771
14. Quarterback (QB)	69	.797	.725
15. Fullback (FB)	69	.812	.797
16. Running Back (RB)	135	.778	.741
Overall	1518	.769	.756

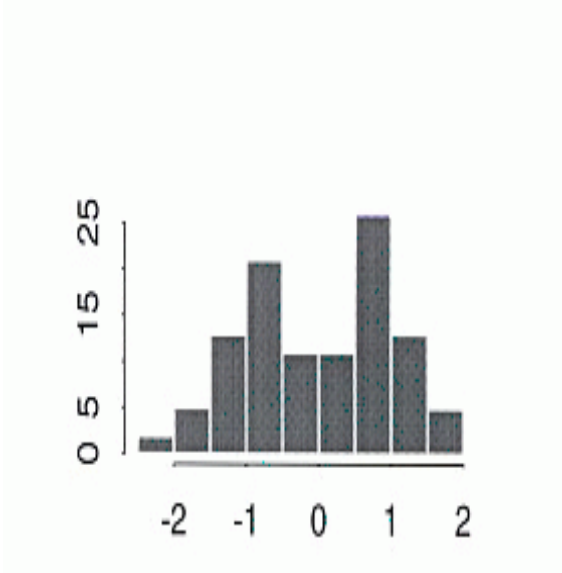
Notes: # = number of players

This tool was initially written in C and Splus by Clive Loader and later modified by Jiayang Sun for classroom use in a hope to combine the best of manual projection pursuit and automatic projection pursuit. A projection pursuit algorithm is automatic if the sequence of projection pursuit solutions is chosen by a computer algorithm. A projection pursuit algorithm is manual if the user chooses by eye which

projections to investigate. For 1-d PP, the dynamic PP provides random projections of the data (nine at a time) displayed as histograms or other graphics for initial examination and then allows user controls for optimization and further searches from these projections. It also calculates P-values of PP and provides additional options. For 2-d PP, Jeremy W. Fleischer under the guidance of Dr Sun has added the features that allow 6 initial random searches and further options and some neat controls.

The coefficients from interesting projections were compared to those from Fisher's linear discriminant rule by position. An orthogonality check found no pairs of orthogonal coefficients. The theoretical analyses for this and a new projection pursuit index inspired from this finding will appear in a different paper.

Figure 3: Projection for Free Safety

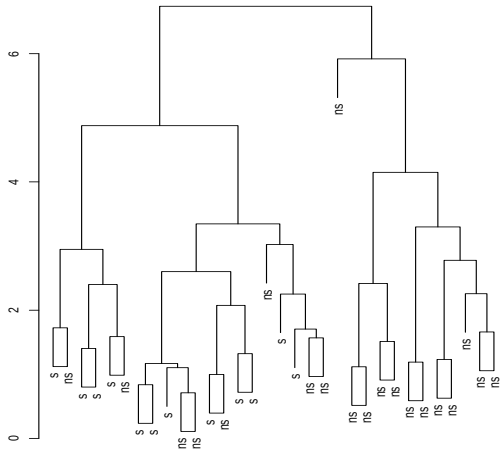


3.3 Cluster Analysis

As a final analysis, an attempt was made to cluster the members of each position into starters and non-starters. The variables used are the same as those used for discrimination, including the imputed missing values. Since the variables are measured on different scales, they were first standardized to a mean of zero and variance of one. Hierarchical cluster Analysis (cf. Everitt and Dunn, 1992) was then performed on this scaled dataset using the Euclidean metric for distance, and the three methods for linkage -connected, complete and average. The resulting clusters were an agglomeration of confused groupings, showing that it is not possible to cluster the players based upon the given set of variables. For most cases, connected and average clustering did not cluster into identifiable groups at all. Complete clustering broke up the players into 2 or 3 clusters though quite inefficiently.

We would like to see two clusters for each position. The fact that we can see more than two clusters may be caused by the situation in which some players are occasionally both starters and non-starters for some positions. The only example that clusters nicely is shown in Figure 4, using complete linkage.

Figure 4: Nose Tackle Clustering



4 Conclusions

In summarizing the results, we can draw some specific conclusions. It's clear that the variables divide into subsets that represent physical characteristics such as lower body strength, upper body strength, running speed, etc. Any information that measures these attributes is useful for imputation. This is true across positions.

Useful discrimination can be found using either Logistic DA or CART. These techniques are more effective for discrimination than projections or linear DA. Clustering was not an effective technique at all. In addition, the authors were not able to find a projection that was orthogonal to direction suggested by Fisher's linear discriminate rule.

We would like to reiterate that the results could be better if we had a more complete data set. Nev-

ertheless, this analysis reveals a successful story of using an imputed data set in a very complex situation. Our classification trees and logistic discriminant rules still provide useful means at the present stage for coaches, athletes and those interested to improve athletes training strategies or build a better football team. Of course, the logistic discriminant rule and CART are doing well for this data and may perform differently in the future since higher and stricter player performance standards are being set each year. It is desirable to see if the imputation provided by the Neural Network technique (the authors are working on it) provides a more efficient and accurate discrimination rule.

The most important interpretation here, which can be applied to each position individually, is that there are additional considerations, not reflected in this dataset, used to determine whether a player will be a starter or not.

Acknowledgments:

The authors would like thanks Mr. Bill Black for his collection of the data and Tom Ryan for his knowledge of football and weight lifting techniques.

References

[1] Berg K., Latin R. W., and Baechle T. (1992). Physical fitness of NCAA Division I football players. *National Strength and Conditioning Association Journal.* 13(3), pp. 68-72.

[2] Breiman, L., J. Friedman, R. Olshen and C. Stone (1984).*Classification and regression trees.* Second Edition. The Wadsworth statistics/probability series.

[3] Craft, J. (1992). Football core exercises of selected universities. *National Strength and Conditioning Association Journal.* 14(2), pp. 34-38.

[4] Everitt, B. and Dunn, G. (1992). *Applied multivariate data analysis.* New York: Oxford University Press.

[5] dos Remedios R., Holland G. (1992). Physical and Performance Characteristics of Community College Football Players. *National Strength and Conditioning Association Journal.* 14(2), pp. 9-12.

[6] Bridgman, R. (1991). A coach's guide to testing for athletic attributes. *National Strength and Conditioning Association Journal.* 13(2), pp. 34-37.

[7] Manski, C. J. (1993). The selection problem in econometrics and statistics. *Econometrics,* edited by G. S. Maddala, C.R. Rao and H. D. Vinod. 73-84. Amsterdam; New York, North-Holland.

[8] Patil, G. P. and Rao, C. R. (1977). The weighted distributions: a survey of their applications. In *Applications of Statistics. Proceedings of the symposium held at Wright State University, Dayton, Ohio, 14-18 June 1976* P. R. Krishnaiah, ed. North-Holland, Amsterdam. 383-405.

[9] Ripley, Brian D. (1996). *Pattern recognition and neural networks.* Cambridge University Press.

[10] SAS Institute Inc. (1994), *SAS/Stat User's Guide Vol. 2, Version 6, Fourth Edition.* Cary, NC.

[11] D. F. Swayne, D. Cook, A. Buja (1998). XGobi: Interactive Dynamic Data Visualization in the X Window System, *Journal of Computational and Graphical Statistics*, **7** (1)

[12] Sun, Jiayang (1998). Projection Pursuit. *Encyclopedia of Statistical Sciences* (updated volumes). Edited by: Samuel Kotz, Campbell Read, David Banks, and Norman Johnson, Vol. 2, pp 554-560, Wiley.

[13] Sun, Jiayang (1998). Good Apples? Sampling Biases. Invited Talk at ASA Cleveland Chapter, May 6, 1998.

[14] Sun, Jiayang and Woodroofe, M. (1997). Semiparametric estimation for biased sampling models. *Statistica Sinica.* **7**, 545-576.

[15] Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, No. 2, 616-620.

[16] Venables, W. N. and B. D. Ripley (1997). *Modern applied statistics with S-Plus.* New York: Springer-Verlag.