9.5

STA571: Advanced machine learning

Classifying Olympic Athletes by Sport (March 4, 2014)

Authors: Matt Dickenson

1 Motivation

To what extent do environmental or biological traits determine sporting success? At the highest level of amateur sports–the Olympic games–we can notice differences in the physical characteristics of participating athletes across sports. For example, the median height for male participants in the 2012 games was 182 centimeters, but the median male basketball player was 202cm while half of male gymnasts were shorter than 167cm. Similar disparities exist in weight and age.

Rephrase conclusions

Can these differences be exploited to classify individuals by sport or event given their physical attributes? If this could be done with a high degree of accuracy, that would suggest that athletes are increasingly optimizing for success by selecting sports for which they are naturally predisposed. If not, we could infer that training makes more of a difference than genetically determined features. To study this question, I will use data on participants in the 2012 Summer Olympic Games published by *The Guardian*.

The way this is phrased makes it sound like an intentional selection strategy. perhaps it is an effect of increased genetic diversity due to more individuals w/ access to the sport as well?

2 Problem definition

There is a pretty famous stat problem about how tall fathers tend to have short sons...etc.

The goal of this project is to predict an athlete's Olympic event given their height, weight, age, and gender. Given data $\mathcal{D}$ on these $p = 4$ features for $n = 9,038$ athletes,[1] our goal is to predict of athletes' events $\mathbf{y}$, which are nested in sports.[2] The quantity we are interested in, then, is $p(\mathbf{y}|\mathcal{D})$.

Figure 1 displays the height, weight, and sex of 487 participants (258 basketball players and 229 weightlifters). Although not all clusters are as distinct as these, this does suggest that clustering methods could accurately classify Olympic athletes by event.

~~Result~~

Can't remember the name but it might be Galton

3 Models and methods

To classify athletes by event, this project will employ Bayesian hierarchical clustering (BHC). BHC models will be trained on subsets of the data using k-fold cross-validation with (randomly sampled) training sets of $10, 20, \ldots, 90$ percent of the original dataset. This method is sensible for Olympic athlete data because participants in two different events within the same sport category should be more similar than those in different sports. For example, we can discover whether an athlete running the 100m event has more in common with a 5000m runner (both distances are classified under "Athletics") or a volleyball player.

[1] There were 10,383 participants in the 2012 games. The 9,038 figure accounts for missingness on the feature variables.

[2] In the Olympic games, "sports" are somewhat broad categories, such as athletics and weightlifting, whereas "events" refers to specific competitions such as the men's 100-meter race.

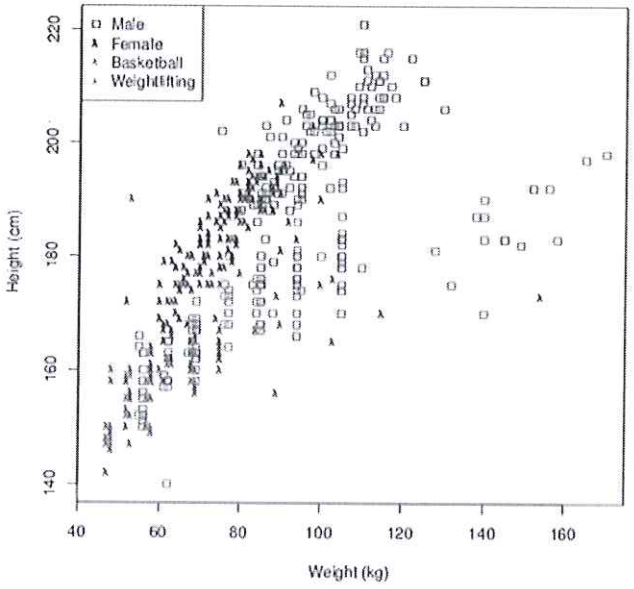

Figure 1: 2012 Olympic Participants in Basketball and Weightlifting, by Height and Weight

In order to classify an athlete from the test set, we can use the posterior predictive distribution

$$p(\mathbf{y}|\mathcal{D}) = \sum_{k \in N} w_k p(\mathbf{y}|\mathcal{D}_k),$$

where w_k is the weight on cluster k, $\mathcal{D}_k$ is the feature data for observations in cluster k, N is the set of nodes in the tree, and N_k represents the nodes lying on the path from the root to node k's parent.

4 Results and validation

The results of this method will be compared to classification with a multinomial GLM and CART. To assess these methods, we will use classification accuracy as measured in the cross-validated models (this is necessary because a single run of BHC or CART does not express uncertainty in the classifications). BHC should do well because it relies on comparing marginal likelihoods of subtrees for merging, rather than distance metrics.

In substantive terms, these results will show whether an athlete's event can be accurately predicted by their physical features. If not, that would suggest that athletic ability can be understood as a latent trait that can be applied to the sport of one's choice. If the model can accurately predict participation in an event using the features specified, that suggests that athletes choose the event that best leverages their natural predisposition. The answer to this question will be important we ponder the future of athletic performance.

okay! this is stated much more cleanly!