Dirichlet Processes for Injury Prevention using Running Activity Categorization

Scott Sikorski
nqj5ak@virginia.edu

1 Introduction

1.1 Background

Long distance running has been an increasingly popular method to promote healthy lifestyles. It has cardiovascular and immune system benefits Ruta et al. [2024], can be a mental health outlet, or provide a source of competition and motivation. However, there can be high risks as 17% - 50% of people will suffer a running-related injury per year Taunton et al. [2002]. It was found that greatly increasing factors such as mileage, time running, and intensity resulted in a higher injury incidence rate Van Gent et al. [2007].

These factors are associated with each run they do and can be used together to find a person's baseline and progression. As a result, Dirichlet Process Models Teh et al. [2010] offer a beneficial approach to labeling runs unique to each person. The growing nature of the clustering allows for training plans that include any number of run types. A person is not confined to what other runners are doing or designating universal terms. Additionally, a DPM approach can allow for natural fitness growth over time by the underlying Gaussian Mixture Model to alter with more data. Or when a run is distinctively different from recent data of what the model or person believes, there exists a possibility of overtraining.

DPMMs offer a flexible and data driven approach to clustering where user's data can drastically differ from each other or over time. DPMMs are modeled as Dirichlet Processes (DP) with a clustering mechanism that provides clusters that follow a probability distribution, Gaussian in this case, and use Gibbs Sampling to infer number of clusters and assignments. The DP is considered a prior for the cluster distribution which, when using the Chinese Restaurant Process (CRP), uses a probability proportional to the number of data points in a cluster or a new cluster given α. Equation 1 shows this where n_k is the number of points in cluster k of K clusters. The clusters that are formed can be modeled as a Gaussian mixture model. In turn, it is determined by Gibbs sampling inference.

$$P(x_i \to k) = \frac{n_k}{\alpha + n - 1}, \quad P(x_i \to K+1) = \frac{\alpha}{\alpha + n - 1} \tag{1}$$

1.2 Relevant Work

Teh Teh [2009] and Orbanz Orbanz and Teh [2010] Orbanz [2012] present introductions to building Bayesian nonparametric models. They focus on Dirichlet Process Mixture Models and present CRP and Stick Breaking as clustering mechanisms. In this work, the CRP forms the basis for the implementation of cluster assignments. Others Rogers et al. [2019], Richardson and Hartman [2018] suggest approaches for using nonparametric clustering for health care settings when regular regression models are insufficient.

Preprint. Under review.

Figure 1: Hierarchical Run Type Representation

2 Methodology

The dataset that I am using is my own running activities from the past year downloaded from Garmin Connect. Each row is a running activity with the main overall stats. I have chosen to focus on Moving Time (in seconds), Distance (in user units), Avg/Max Heart Rate, Avg/Max Speed, Avg/Max Power. These 8 features have distinct correlations to the type of run that I'm attempting to categorize.

For preprocessing, I have chosen to normalize the data using the min-max normalization with a feature range of $(0, 1)$. This ensured that the data would give equal weight to each attribute. Principal Component Analysis Abdi and Williams [2010] was then used to reduce the dimensionality to 3d. In initial testing without preprocess PCA, the number of clusters would blow up to the max number. With PCA and min-max normalization, the DPMM more accurately modeled the run clusters with 8, as seen in section 3. I found all of these preprocessing steps extremely necessary as it presents a more approach to the features. At first, figure 5.3 showed a large spread in the most contributing factors. Time was measured in seconds and would sometimes be 2000x larger than the maximum heart rate.

As referenced earlier, the core algorithm of the clustering focuses on performing iterations of cluster assignments until convergence is met. For each data point, the prior is calculated of the existing clusters. The posterior can then be found from this prior and the likelihood of each cluster from its Gaussian probability distribution. The new cluster probability is found given hyperparameter tuning of the prior mean and variance. The data point can then be assigned to the most likely cluster.

2.1 Evaluation

To evaluate the run cluster categorization, I have labeled each run. This falls under Training Run, Workout, Long Run, Warmup/Cooldown, Recovery, Shakeout, Double, Race. See section 5.1 for extra info on these groups and distinctions and Figure 5.1 for the dataset of these. Once initial clustering is done, the cluster number can be mapped to each run type. Adjusted Rand Index (ARI) Santos and Embrechts [2009] is used to assess the accuracy and similarity of the learned assignments. I chose this metric as to judge the raw clustering ability without needing all run types matched. This is because some people may choose to not include certain run types with their training plan.

Additionally, I will evaluate the predictive performance of new data samples. The user will input their run for the day with their self-evaluation of what they planned. The model will determine its group and see if it matches. If not, the mismatch could indicate the possibility of overtraining, using a hierarchical representation (Figure 2.1). This classification task will use standard accuracy and f1 score metrics to analyze the performance. An expansive data partition is needed to test each type.

3 Preliminary Results

I started the process by using SkLearn's approximate implementation, specifically the Bayesian Gaussian Mixture Xue and Guillemont [2024]. It uses the Dirichlet process for weight concentration

prior type and the α hyperparameter for weight concentration prior weight. This implementation found 7 clusters which perfectly matches to the number of distinct annotated labels. Figure 5.3 shows that the trends that time leans to -1 PC1 with max pace, heart rate, and power being the higher factors of PC2 = +1. This categorizes well to the proportions of labeled runs. I will present a more concrete number with my implementation and the evaluation.

Additionally, figure 5.3 provides more insight to how close clusters can differ in PC3. The few data points in cluster 5 are extremely long and strenuous workouts that verge on long distance race efforts. These showcase how a large departure from the recent data can indicate a possible new effort that extremely taxes the runner. With this, the classification algorithm can recommend an easier day to ensure proper recovery.

In similar fashion, the clustering algorithm is at times prone to classifying some runs close to each other. A double, typically 4-6 miles at training run pace or a little shorter, can easily be confused with a shorter recovery run that made be done quicker due to the components being extremely close. However, this is where the hierarchical representation will benefit from a more triangle structure. Runs at the bottom, double, recovery, or training runs, can be intermixed due to their similar nature. Once the runner reaches the next up layer, it will be a warning sign.

4 Future Work

1. Gather results using my DPMM implementation. All results are using the Sklearn implementation.
2. Do extra testing on including more run features and weighting the features differently. I have up to 4 features including elevation gain and Garmin effort metrics to use. As well, I would weigh existing and new features after performing the min-max normalization. I hypothesize that average pace and average heart rate would benefit from being weighed more heavily.
3. Perform classification task on test set to analyze overtraining. This involves partitioning the data, use learned model parameters to predict, compare to true label, and adding the sample to the DPMM.

   Stretch Goals
4. Augment the dataset and test with other person's data. The overall goal should be that each person develops their own DPMM to assess their training. My data has all types of possible runs, but I would like to test with data excluding some labels.
5. Try to have more specific labels on some run types. Through current testing, the model cannot detect subtle differences in workout types that can be focused on aerobic or anaerobic benefit. However, this may not be relevant for most people who are not training at high levels.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Peter Orbanz. Lecture notes on bayesian nonparametrics. *Journal of Mathematical Psychology*, 56: 1–12, 2012.

Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. *Encyclopedia of machine learning*, 1:81–89, 2010.

Robert Richardson and Brian Hartman. Bayesian nonparametric regression models for modeling and predicting healthcare claims. *Insurance: Mathematics and Economics*, 83:1–8, 2018. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2018.06.002. URL `https://www.sciencedirect.com/science/article/pii/S0167668717305437`.

T.J. Rogers, K. Worden, R. Fuentes, N. Dervilis, U.T. Tygesen, and E.J. Cross. A bayesian nonparametric clustering approach for semi-supervised structural health monitoring. *Mechanical Systems and Signal Processing*, 119:100–119, 2019. ISSN 0888-3270. doi: https://doi.org/

10.1016/j.ymssp.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S088832701830623X.

Damian Ruta, Bogumił Bocianiak, Anna Kajka, Julia Hamerska, Joanna Antczak, Laura Hamerska, Urszula Fenrych, Karolina Wojtczak, Olga Skupińska, and Julia Lipska. Health aspects of amateur long-distance running. *Quality in Sport*, 20:53342–53342, 2024.

Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer, 2009.

Jack E Taunton, Michael B Ryan, DB Clement, Donald C McKenzie, D Robert Lloyd-Smith, and Bruno D Zumbo. A retrospective case-control analysis of 2002 running injuries. *British journal of sports medicine*, 36(2):95–101, 2002.

Yee Whye Teh. An introduction to bayesian nonparametric modelling. *Machine Learning Summer School*, 2009.

Yee Whye Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

RN Van Gent, Danny Siem, Marienke van Middelkoop, AG Van Os, SMA Bierma-Zeinstra, and BW Koes. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *British journal of sports medicine*, 41(8):469–480, 2007.

W. Xue and T. Guillemont. Scikit-learn: Bayesian gaussian mixture, 2024. URL https://scikit-learn.org/1.5/modules/generated/sklearn.mixture.BayesianGaussianMixture.html.

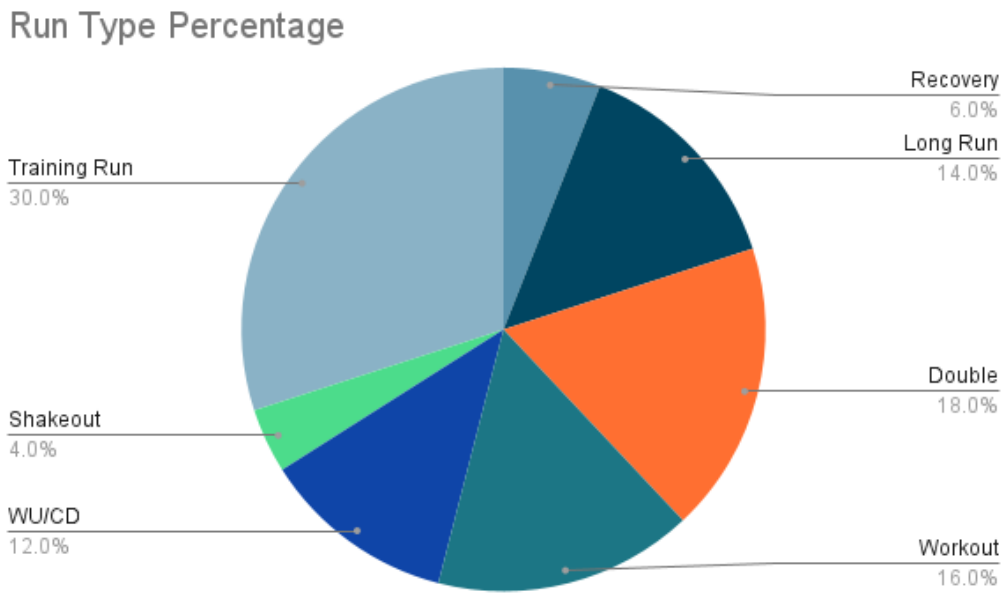

Figure 2: Dataset Makeup

5 Supplementary Material

5.1 Run Type Distinctions

Long Run (LR): About 1.5-2x longer than the TR with a slightly elevated pace.

Workout (WU): Dedicated sessions where average and max pace, heart rate, and power are elevated but not spending as much time in those stats as a race.

Race: The most extreme case where average heart pace, pace, and power are close to their max counterparts. These averages will also be higher than any other run that is done.

Warmup/Cooldown (WU/CD): These are short 15-20 minutes jogs done at TR pace prior to workouts or races with low or high heart rates depending on intensity intention.

Shakeout (SO): Less than 2 miles at a slow pace just to wake up the body.

Double: Secondary run done on the same day as another run. Variable heart rate and pace depending on how one feels. Generally 25-35 minutes. Most unlikely run for beginners to do

Recovery: Usually shorter time/distance but does not depend on it. Intention on low heart rate and low power output

Training Run (TR): The most common run and general running. Variable feature values but are not high or low as the other runs with emphasis on being the middle of all features.

5.2 Source Code

Source code can be found at `https://github.com/sgsikorski/STAT6020/tree/main`. The src folder contains the effective source code with my written DPMM in src/DPMM.py. Any output lies in res. The cluster plots in 2d and 3d are found along with the parallel coordinates plot. See the current results below.

5.3 Result Plots

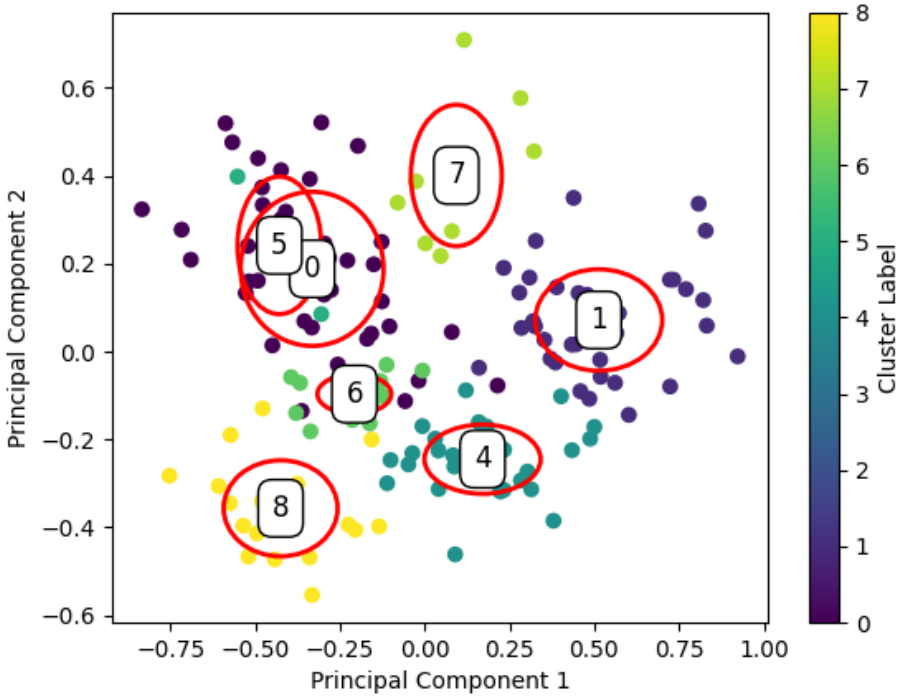

Figure 3: 2d Cluster Representation

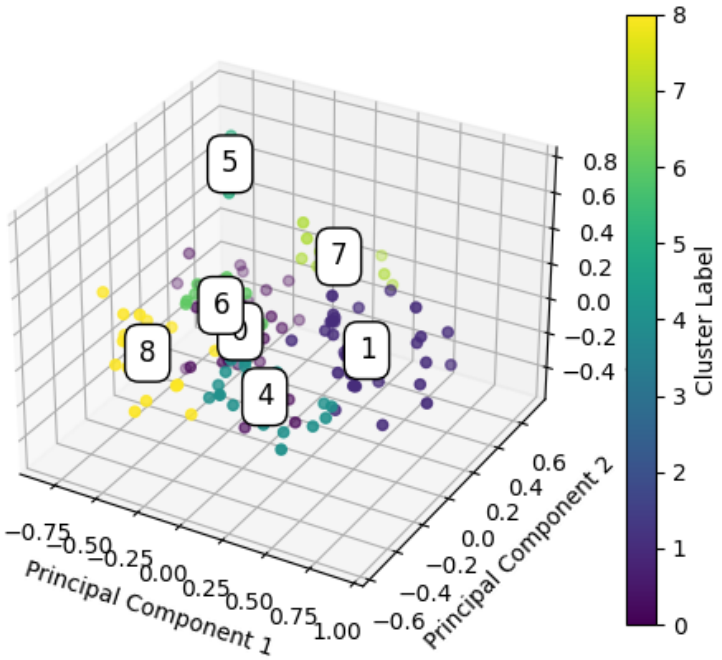

Figure 4: 3d Cluster Representation

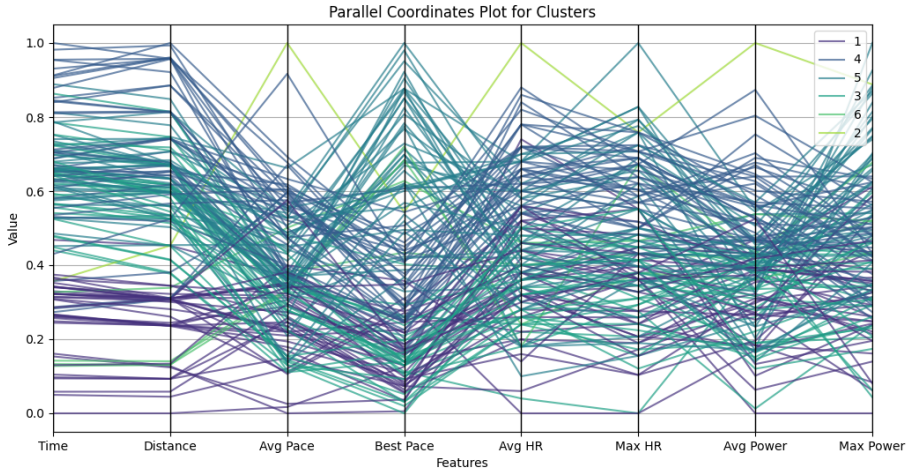

Figure 5: Parallel Coordinates of Features and Their Clusters