"Walk or Run" Report

---

1. **A Descriptive analysis**

   (a)Please describe the format of the data files.

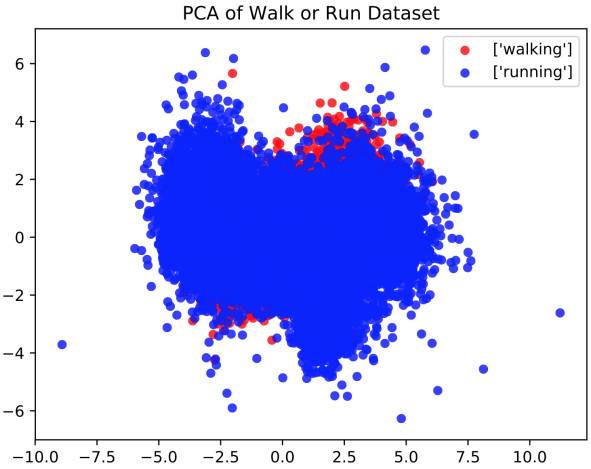
   (b)Can you identify any limitations or distortions of the data?

   (a)

   This dataset contains 37776 samples and each sample has 6 different features. If we convert it into a 2D matrix, it will be a 37776 by 6 one. If we want to plot it directly, it will be in 6 dimensions.

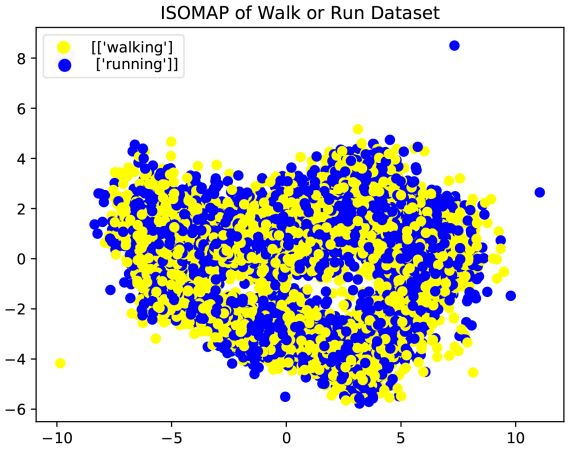
   Method1: PCA

   Since the data points are near in the 6D space, and their euclidean distance are short, the plot has a significant overlap on a 2D plot.



   Method2: ISOMAP

   Form method 1, we notice that the data in high space are really near in euclidean distance, so we use geodesic distance and k nearest neighbor which is ISOMAP. Here we use 10 nearest neighbor.



This result is kind of better that the result from PCA, but there is no clear boundary between two classes(walking and running), so this means that the contribution of all 6 features to the result are small. If we want a more clear visualization, we may drop some of the features. However, the features can be conditional independence, so there are more trials needed to do the feature selection in order to have a better visualization.

(b)

There can be some limitations of this dataset. We can find that data are collected form only one user, and this can be a problem if we want to make a general decision the a person is walking or running.

2. **B Model building**

Method1: KNN

CNN training error: 0.470635347336
CNN test error: 0.468645217059
number of variables in subset: 738

This solution has a huge error. From part A, we already know it is difficult to find a boundary for this dataset in a 2D space, so this result is kind of reasonable that KNN not works well.

Method2: Neural Network

Training error = 0.006935620499788242
Test error = 0.010853451927149482

Here we use MLPClassifier to do the analysis. There is one hidden layer and 100 hidden units. Since the data size is not so big and it contains information from only one user, too many hidden layers will lead to overfitting. A large number of hidden units provide more neurons in neural network which can increase the test accuracy. Overall, the MLPClassifier works great here, and the results is quite satisfied.