CS434 Final Project Report

Monica Sek and Arthur Shing

1 Feature formulation and preprocessing

1.1 Features

The features we are feeding our learning algorithm are time (specifically the hour), glucose, slope, IOB, MOB, morning, afternoon, evening, night, and hypo in 30 mins. We flatten the rows based on their columns, so for every 7 rows all the values of the first column would be compiled then is followed by all the values of the next column. This was all generated into a numpy matrix. To make sure we accounted for positive instances, we scanned through for a positive hypo and took the 6 prior rows and flatten them. To account for negative instances, we scanned through the rest of the data in rows of seven, if it happened to be positive it will discard that instance otherwise it will keep it.

1.2 Preprocessing

We normalized our data so that any non-binary values will become binary. This allowed to reduce the noise going into our algorithm and provide and more accurate prediction.

2 Learning algorithms

2.1 Algorithms explored

We chose the following algorithms for this project: KNN, logistic regression with batch gradient descent, and linear regression

First, we chose KNN because for positive outcomes, it will grab the nearest neighbors and be able to find similar instances.

Next, we chose logistic regression with batch gradient descent because it computes fewer updates by only updating after the end of a training epoch. It also has the tendency to result in a stable convergence.

Finally, we chose linear regularization because it will generate a general line of best fit, that incorporates multiple of features.

2.2 Final models

The final model that we chose is KNN, where K=20.

3 Parameter Tuning and Model Selection

3.1 Parameter Tuning

3.1.1 KNN

We changed the following parameter: K to grab a certain amount of neighbors near an instance.

3.1.2 Logistic Regressions with Batch Gradient Descent

We changed the following parameters: learning rate to control how fast it learns, the epochs to tell it how many times to run (this is to update), and the prediction rounding to assist our predicted value to result in the value we believed.

3.1.3 Linear Regression

We changed the following parameters: N/A

3.2 Model selection

We chose to use KNN in the end because it provided us the most accurate dataset compared to logistic and linear regression.

4 Results

The algorithm we used weren't the best option to create models with the given set of data. KNN didn't often have a positive instance near them because they were so scarce.