Type I diabetes: Hypoglycemia Prediction

Final project

Background

- Type I diabetes patients cannot insulin, a hormone that allows the body to absorb glucose (sugar) in the blood and turn it into energy for consumption
- They have to constantly monitor their blood sugar level and administer insulin regularly to keep their blood sugar levels under control
- Overly high blood sugar (Hyperglycemia): extra burden on your organs, potential damage to long term health
- Overly low blood sugar (Hypoglycemia): body might start shutting down and consequence could be lethal
- We are interested in predicting hypoglycemia for type I diabetes patients

Data

Time	Glucose	Slope	IOB	MOB	Morning	Afternoon	Evening	night	hypo in 30m?
2016-04-05T03:34:46Z	79	10.5	0.073333	0	0	1	0	0	0
2016-04-05T03:39:46Z	88	10	0.073333	0	0	1	0	0	0
2016-04-05T03:44:46Z	101	11	0.073333	0	0	1	0	0	0
2016-04-05T03:49:46Z	106	9	0.073333	0	0	1	0	0	0
2016-04-05T03:54:46Z	113	6	0.073333	0	0	1	0	0	0
2016-04-05T03:59:46Z	122	8	0.073333	0	0	1	0	0	0
2016-04-05T04:04:46Z	132	9.5	0.073333	0	0	1	0	0	0
2016-04-05T04:09:46Z	142	10	0.073333	0	0	1	0	0	0
2016-04-05T04:14:46Z	149	8.5	0.073333	0	0	1	0	0	0

Time: time of measurement (in UTC)

Glucose: the glucose reading

Slope: the slot of the glucose curve

Hypo in 30m?: Our target variable, indicating if a hypo event happens within 30 minutes

IOB: insulin

MOB: meal on board

Both quantities are computed based on intake and some physics model about how human body metabolize inputs like insulin and food/carbs

A closer look into the data

Time	Glucose	Slope	IOB	MOB	Morning	Afternoon	Evening	night	hypo in 30m?
2016-04-21T01:58:59Z	96	-2.28E-14	0.073333	4.22E-18	0	1	0	0	0
2016-04-21T02:03:59Z	96	1	0.073333	2.34E-18	0	1	0	0	0
2016-04-21T02:09:01Z	95	-0.5	0.073333	1.29E-18	0	1	0	0	0
2016-04-21T02:13:59Z	96	-2.28E-14	3.5733	7.16E-19	0	1	0	0	0
2016-04-21T02:18:59Z	91	-2	0.073333	3.96E-19	0	1	0	0	1
2016-04-21T04:28:59Z	74	4.5	0.073333	8.30E-26	0	1	0	0	0
2016-04-21T04:33:59Z	77	3.5	0.073333	4.59E-26	0	1	0	0	0
2016-04-21T04:38:59Z	80	3	0.073333	2.54E-26	0	1	0	0	0
2016-04-21T04:43:59Z	80	1.5	0.073333	1.41E-26	0	1	0	0	0

- This measurement is at 2016-04-21T02:18:59Z, its label is 1, which means 30 minutes later, a hypo event happened (2016-04-21T02:43:59Z)
- This hypo event lasted for almost 2 hours
- All the time points after 2016-04-21T02:18:59Z, until the event is over, the data is removed
- Our data resumes at 2016-04-21T04:28:59Z, when the subject’s glucose returned to normal

Goal

Given a 30-minute window of data, predict whether the patient will have a hypoglycemia event within 30 minutes

Input (x)

Time	Glucose	Slope	IOB	MOB	Morning	Afternoon	Evening	night	hypo in 30m?
2016-04-05T03:34:46Z	79	10.5	0.073333	0	0	1	0	0	0
2016-04-05T03:39:46Z	88	10	0.073333	0	0	1	0	0	0
2016-04-05T03:44:46Z	101	11	0.073333	0	0	1	0	0	0
2016-04-05T03:49:46Z	106	9	0.073333	0	0	1	0	0	0
2016-04-05T03:54:46Z	113	6	0.073333	0	0	1	0	0	0
2016-04-05T03:59:46Z	122	8	0.073333	0	0	1	0	0	0
2016-04-05T04:04:46Z	132	9.5	0.073333	0	0	1	0	0	0

Want to predict (y)

Unpacking the data into instances

Time	Glucose	Slope	IOB	MOB	Morning	Afternoon	Evening	night	hypo in 30m?
2016-04-21T01:58:59Z	96	-2.28E-14	0.073333	4.22E-18	0	1	0	0	0
2016-04-21T02:03:59Z	96	1	0.073333	2.34E-18	0	1	0	0	0
2016-04-21T02:09:01Z	95	-0.5	0.073333	1.29E-18	0	1	0	0	0
2016-04-21T02:13:59Z	96	-2.28E-14	3.5733	7.16E-19	0	1	0	0	0
2016-04-21T02:18:59Z	91	-2	0.073333	3.96E-19	0	1	0	0	1
2016-04-21T04:28:59Z	74	4.5	0.073333	8.30E-26	0	1	0	0	0
2016-04-21T04:33:59Z	77	3.5	0.073333	4.59E-26	0	1	0	0	0
2016-04-21T04:38:59Z	80	3	0.073333	2.54E-26	0	1	0	0	0
2016-04-21T04:43:59Z	80	1.5	0.073333	1.41E-26	0	1	0	0	0

- Take each instance and break it into continuous chunks.
- Each chunk will be ending with one positive time point
- Generate one positive instance for each chunk by taking the 30 minute time window
- Other 30-minute windows in the chunks are all negative instances
- In general you will get a lot more negative instances than the positive ones

Set up of the problem

- We have data from 10 subjects. One subject has too few Hypo events and is removed from consideration for this project
- It is not clear whether we can expect generalization across subjects, so we will test this out.
- We will set up the training/testing in two different ways
  - Within subject: We will provide training data for two subjects, so that you can build a predictive model for each subject. We will provide test data for each subject for you to submit your prediction, which will be evaluated against the ground truth
  - Across subject: We will provide the complete data for another set of 4 subjects and hold out the remaining 3 subjects for testing and evaluation.

Evaluation criterion

- An important fact: very limited number of positive events
- On average, we only have 30-40 positive events per subject for 6000 – 7000 time points
  - Highly unbalanced
- By always predicting negative, we will get over 99% accuracy
- Evaluation criteria to consider in this context:
  - Precision, recall and F1 measure of the rare class – for binary predictions
  - Area under the ROC curve - for probabilistic predictions
  - Both criteria will be introduced in the outlier detection lecture

Set up of the competition

For training:

- You will be given training data for each of the tasks
  - General population prediction task: you will have four separate files, each containing data from one of the four training individuals. You will need to **aggregate these data** to train **one model for the general population** that will be tested on three remaining individuals during testing/scoring
  - Individualized prediction task: you will be given two separate files, one for each individual. You will train **a model for each individual**, each will be tested on its corresponding holdout test data during testing/scoring

Set up of the competition

Test data

- For each of the three models you train, you will be given a test set.
  - General-test for the general population model
  - Individual-test-1 for individual 1
  - Individual-test-2 for individual 2
- The format of the test data will be slightly different from the training data, which contains continuous time series

- We will break the time series data into individual instances, with each instance described by a separate csv file contains 7 rows of data (corresponding to data from a 30-minute window).

- Sample test files will be given to you to test your program.

Set up of the competition

- Submission
    - For each test set, you will need to submit a prediction file in csv format. One line for each test instance. In each line, you will provide two values.
        - The first is a numeric value, either a probability or a continuous-valued score, to indicate how likely that instance will see a hypo event in 30 minutes (higher value indicate higher likelihood)
        - The second value is a binary prediction, 1 for hypo event, 0 for not hypo event
    - You are required to explore at least three different methods and produce three different prediction files for each test set
    - You will also be required to submit a completed questionnaire to describe your efforts for this assignment, which will be considered during grading

Set up of the competition

- Competition
  - For each task, we will score all the submitted predictions
    - For binary prediction, we will evaluate the predictions using the F1 measure of the rare class
    - For the continuous score, we will evaluate the predictions using the Area under the ROC curves
  - Each team is required to submit exactly three prediction files for each task
  - Your team will be scored by your best entry

Some things to consider

Disclaimer: I have done quite a bit in preparing the data, but have not build any model on them yet

- Handling the features
  - Each data instance is described by the glucose readings, and other measurements or computed quantities from a 30-minute time window.
    - The easiest thing would be to simply ignore the temporal structure in the features, and treat it as a flat one and let the learning algorithm figure it out
    - Alternatively, one could just consider the last time point and ignore the previous ones
    - More sophisticated approaches?
      - Use PCA to reduce the features and remove redundancy?
      - Use CNN or RNN to learn better representation of the features
      - …

Somethings to consider

- Approach
    - First you need to unpack the time series data and generate training instances
    - A critical issue: very unbalanced class distribution
        - The training instances are not generated yet – they are still all packed into time series. When you generate the instances, you can potentially apply some subsampling to the normal instances and remove some redundant instances (many of the normal instances overlap in time)
        - You can introduce weights to give higher weights to hypo examples
    - What algorithms should you try?
        - This is a classic classification task. Many methods can potentially be tried
    - How to set up validation for parameter tuning and model selection?
        - For general population task: does it make more sense to do leave one subject out validation? Or leave portion out validation?
        - For individual model task, how can you structure your folds so that you are not including very similar instances in training and validation?