DS-GA 1006 Final Report
Trend Modeling in Childhood Obesity Prediction

Fu Shang FS1520@NYU.EDU
NYU Center of Data Science

Ethan Wu YW3375@NYU.EDU
NYU Center of Data Science

Yi Zhou YZ4525@NYU.EDU
NYU Center of Data Science

Daoyang Shan DS5471@NYU.EDU
NYU Center of Data Science

Abstract

Childhood obesity has become an increasingly obvious phenomenon in the United States. Researchers in NYU Langone Health have conducted researches in predicting potential childhood obesity in the early stage of growth by utilizing machine learning tools. Based on their work, our team performed further exploration, especially in children growth trend modeling, which we believe could introduce new insights on this topic, using clustering, time series and Neural Networks.

1. Introduction

1.1 Background

In the United States, childhood obesity has been an increasingly obvious phenomenon since the 1970s (Alston and Okrent 2017). In 2016, 18.5% of US population aged from 2 to 19 had obesity with a significantly higher susceptibility for boys than girls (Skinner et al. 2018). Also, recent studies suggest that childhood obesity is becoming increasingly prevalent: data from 2015-2016 showed obesity rates have increased across children of all ages, including a large increment at the youngest ages (2-5 year-old) (Skinner et al. 2018). Due to the strong link between childhood obesity and adult comorbidities, and the challenges in reducing body mass index (BMI) later in life, effective strategies are needed to address the condition early in life. If we were able to predict, as early as possible, the group of children who have higher risk of future obesity, then we would be able to better allocate resources to those most at risk (Hammond et al. 2018).

To start with, we need to closely examine the obesity development and how do we formulate the prediction problem based on that. As Figure 1 shows, ”two critical periods in the development of obesity include the prenatal and infancy period, and early childhood" (Hammond et al. 2018). The first two years mark the first critical period in the development of obesity, and the second period starts at five years of age, where the adiposity rebound marks little change in BMI and a transition into childhood growth. As we can notice, the condition of children at the second period (adiposity rebound) is a good indicator for their adulthood obesity (Hammond et al. 2018). If we can predict such condition of children around their age of roughly 4 to 7 using available data during the even earlier stages of life (for example, the first period, as we discussed above), we are confident to claim that our prediction can also predict their adulthood obesity.

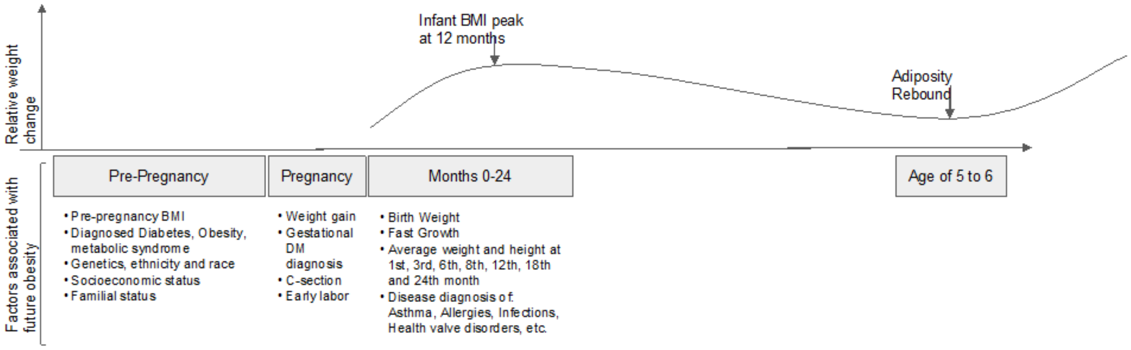

Relative weight change
Infant BMI peak at 12 months
Adiposity Rebound
Factors associated with future obesity
Pre-Pregnancy
Pregnancy
Months 0-24
Age of 5 to 6
• Pre-pregnancy BMI
• Diagnosed Diabetes, Obesity, metabolic syndrome
• Genetics, ethnicity and race
• Socioeconomic status
• Familial status
• Weight gain
• Gestational DM diagnosis
• C-section
• Early labor
• Birth Weight
• Fast Growth
• Average weight and height at 1st, 3rd, 6th, 8th, 12th, 18th and 24th month
• Disease diagnosis of: Asthma, Allergies, Infections, Health valve disorders, etc.

Figure 1: Factors at the prenatal and infancy periods related to childhood obesity

1.2 An Overview of Data Sets

In our project, we used raw data from EHR systems from a secured online health system in New York City. The raw data sets used in this study spans from January 1, 2008 to August 31, 2016 and contained the history of 52,945 children at different ages, "and of 36,244 of their respective mothers for visits ranging from well-child visits to inpatient and outpatient services" (Hammond et al. 2018). Furthermore, other publicly available data, including census and geographical data that describes neighborhood environments in NYC are also available. The volume of raw data, however, makes it hard and unnecessary to go through in detail, and we will discuss the cleaned and re-organized data sets in the next section.

2. Previous Works

Many researches have been done to explore possible machine learning models that are useful for childhood obesity prediction, among which is the research conducted by researchers in NYU Langone Health, including Robert Hammond, Silvia Curado and Narges Razavian. Their work leads to our capstone project, and therefore it is necessary to briefly go through their methods and results.

2.1 Utilization of Existing Data

Previous research at NYU Langone Health has found all children with at least one BMI measurement between the ages of 4.5 and 5.5 years and with at least one visit in the first two years of life. Then, for all children that considered as data points in this research, they confirmed the BMI measured was in the valid range of 10-40 kg/m^2. Hence, the first step is to build features from the filtered data. For maternal data they used vital signs, diagnosis codes, procedure, and laboratory results during six separate time periods ranging from pre-pregnancy to post-delivery (Hammond et al. 2018). For child data, they created features that group vital signs into averages over 11 time periods/points: at birth, 0-1 months, 1-3 months, 3-5 months, 5-7 months, 7-10 months, 10-13 months, 13-16 months, 16-19 months, 19-24 months, and latest reading available (before 24 months), to capture the time frames surrounding the standard well-child visits during the first two years of life (Hagan et al. 2007). They also calculated the change between each of these time periods as well as the change from birth to two years for all vital signs. For all other EHR data, they only created a single feature for each attribute (Hammond et al. 2018).

As the last step, they combined cleaned and filtered EHR data with geographical data from 2015 American Community Survey 5-year Estimates, which gave them description of the neighborhood where those children and parents live (such as education level, unemployment rate, average household income, etc.). Also, they created binary encoding using disease codes from International Classification of Diseases to indicate the presence of different diseases among both children and mothers (Hammond et al. 2018).

With all the methods mentioned above, they created 19,290 attributes for the final data set. However, most of the attributes show a considerable extent of sparsity. In fact, only 12% of all attributes actually contain information, and only 8% among all appears on at least five children (Hammond et al. 2018). Still, such method generates well-organized data set for machine learning model, and we basically inherit from this feature engineering infrastructure. With some additional feature engineering, which will be discussed later, this is essentially the data set we used for our own work.

2.2 Modeling and Result

Using the cleaned data set, the goal is to predict obesity around the ages of 4.5 to 5.5. Formally speaking, the median BMI value for a child between the ages of 4.5 and 5.5 years is calculated, and classified into obese or not. Researchers implemented both regression-based and classification-based models. In the classification task, they used class probabilities to predict the binary outcome of obesity status (obese or not). In the regression task first they normalized the median BMI value and, using the predicted normalized BMI, they classified children as of obese if their predicted values are greater than the threshold for obesity (Hammond et al. 2018).

Various models were applied here and AUC was used as the evaluation criteria. LASSO Regression led to the best performance, with AUC score equal to 0.82 for girls and 0.76 for boys. Besides, LASSO Regression handled sparsity pretty well, utilizing only 35 features for girls models and 144 features for boys model (Hammond et al. 2018). One thing needs to be noticed is that almost all of those important features stay in the children vital sign category, which is fairly reasonable.

Girls LASSO Model	Boys LASSO Model
Maternal Vital: Wt-postPregnancy	Vital: BMI-gain0to24
Vital: BMI-avg16to19	Vital: BMI-latest
Vital: BMI-avg19to24	Vital: Wt-avg19to24
Vital: BMI-avg13to16	Vital: Wt for Length ZScore-latest
Vital: BMI-latest	Vital: Wt-gain0to24*

Table 1: Top 5 Important Attributes for LASSO

3. Performance Boosting on Classification Problem

As the starting point of our project, we attempted to explore further on the original classification problem and get better results. We built SVM, Random Forests, XGBoost and LightGBM Classifiers to predict the target of obesity. Overall, among which LightGBM brought us the best result (Microsoft 2017).

To be more specific, we defined data loader and use validation accuracy as criteria for early termination. With hyperparameter tuning and early termination to prevent overfitting, our best model has an AUC (ROC) of 0.86 for girls and of 0.80 for boys, as shown in Figure 2.

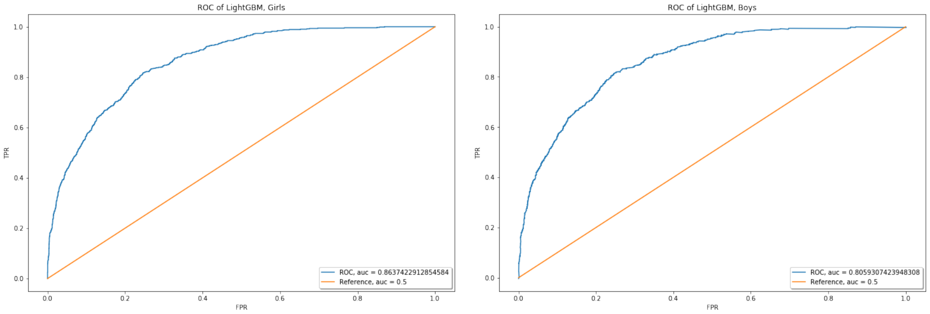


Figure 2: ROC Curves for Best LightGBM

The model we trained beats the baseline model from NYU Langone Health in terms of both AUC and accuracy. The main reason is that LightGBM is a newly developed machine learning algorithm and it deals with medium-sized data well in particular (Ke et al. 2017).

4. Clustering and Trend Modeling

4.1 Problem Formulation

Despite the successful attempt mentioned above, new questions toward this topic arose here. From the perspective of real-world application, we are always interested in not only the obesity condition of an individual child at a certain time point– instead, we hope to know a ”obesity trajectory”: for example, the obesity conditions for a consecutive set of time stamps along the growth of children. On the other hand, in the previous work, data sets were used in an ”discrete way”: each children is treated individually, and all attributes, especially those associated with time stamps, were also utilized individually. Therefore, it may be useful to group all children into different categories, and discover a ”growth trend” for each category using the existing data in a ”time-series-like” manner. These two ideas above, if implemented properly, may bring new insights into this study.

Generally speaking, we aim to explore three areas:

- Group all children into different groups, and see if we can find signature characteristics and fit different growth trajectories, for each group.
- Using the existing data as ”trend of growth”, can we predict obesity condition at the ages of 4.5-5.5 (original goal)
- Furthermore, can we use such ”trend of growth” to predict the ”next trend of growth”? In particular, given the growth trajectory of a child or a group a children from 0-2 year-old, can we estimate the future trajectory?

In the next several sections we will explain the detailed formulation, methods and results of the three areas above.

4.2 Data Exploration

First of all, we tried to explore if we can find the ”natural categories and trends” embedded within the original data. Indeed, one question should be asked: what is the ”original data”? In the previous sections we explained in detail how previous researchers cleaned and integrated data, and the data sets we used in our project are the directly generated by the existing data cleaning infrastructure, with only one distinction: in addition to average vital sign data in 9 time periods (0-2 year-old), we summarized vital sign data on 6 extra periods: 24-27, 27-33, 33-42, 42-54, 54-66 and 66-78 months, and added those averages as extra attributes to the data set we use. The major reason of doing this is because we expected ”series-related” findings on a longer time period, not just 0-2 year-old period.

After adding those new features, we extracted two data sets. For the first one, we extracted all and only vital sign related attributes from 0-2 year-old period, since, according to the previous study, almost all attributes that truly contributed to prediction are vital sign attributes. This data set contains 4970 children in total.

For the second one, it also only contains vital measurements, but we have used the 15-period enriched data introduced in the previous paragraph since we are interested in trend detection as well. The data has 11494 children who have visited the hospital at least once during that age period, so naturally there are much more children than that of the first data set. The obese/non-obese ratio is 2278/9216 for this second data. Specifically, we have focused on BMI Percentile in this data, since BMI balances the effects of both height and weight well, and its percentile is a unitless quantity. For each child, there are 15 data points that record his/her average BMI percentile at the according age bucket. Therefore, we consider the data has 15 features and every entry is a percentile record.

4.2.1 Clustering: Do We Have Categories in Nature?

A quite intuitive thought when we come to ”separating things into categories” is to use clustering analysis, especially when we do not have a strict criterion on determining the performance of such categorization (for example, a traditional classification task with loss function). Indeed, we do need an evaluation standard for clustering, and for our project, an naive but still reasonable goal is that our clustering, regardless the number of clusters we get, can somehow separate obese and non-obese children. Be aware the difference of this goal and a classification problem, for we do not specify the number of clusters (categories) we should have, neither do we enforce a strict definition of ”desirable separation”. To be specific, what we really focused on here was the ”in-cluster obese/non-obese ratio”. Indeed we hoped to see each cluster contains only obese or non-obese children, but we were also interested in if some clusters showed dramatically different ratios compared to those of other clusters.

Before we tried clustering on those two data sets, we plotted the distribution of obese and non-obese children onto two-dimensional plain and hoped to see if those two types of children showed distinct distribution pattern. We applied two dimension reduction methods for visualization: Principle Component Analysis (PCA) and T-SNE. We also normalized those two data sets using Gaussian Normalization.

As we can see from Figure 3 and Figure 4 on next page, obese children (yellow dots) and non-obese children (purple dots) seem to be nested together under both dimension reduction techniques. Of course, we cannot simply conclude that it is impossible to separate those two types of children using clustering, although we may not be able to determine a ”good clustering” visually.

Now we proceeded to the actual clustering. We applied two types of clustering methods: K-Means Clustering (in a broader sense, general Gaussian Mixture Models)

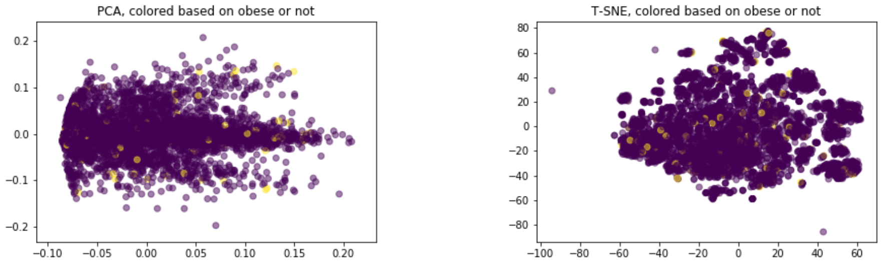


Figure 3: Distribution of Data Set 1

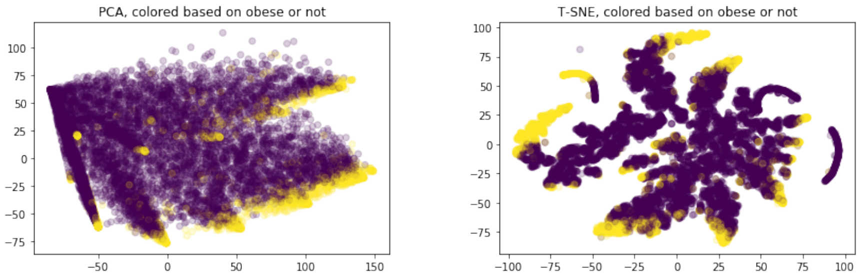


Figure 4: Distribution of Data Set 2

and Hierarchical Clustering (using the Agglomerative Clustering implementation in sklearn). For Hierarchical Clustering, we used the default ward linkage in sklearn, as we noticed that other linkages such as complete and average could not effectively generate clusters in reasonable sizes (almost all points were placed in one cluster). We tried different cluster numbers ranging from 2 to 10, and examined the in-cluster obese/non-obese ratio for each clustering setting. The results of some clustering settings can be found in Table 2 on next page.
Unfortunately, we could not see a huge difference among the in-cluster obese/non-obese ratios of clusters, and obese and non obese children seem to distribute evenly in all clusters. This result might suggest that the "embedded categories", categorized by obesity, among all children may not exist at all.

4.2.2 Categories and Trends

Despite an unclear clustering schema for both data sets, we would like to know whether growth trajectories over time differ among clusters. The differences may not necessarily arise from obesity status, but from other medical conditions.

Below shows an example of the trajectory curves generated from 2-cluster result through hierarchical clustering on the second data set. The curve was created using the Kalman smoothing method (see Appendix). The x-axis consists of the 15-period age buckets. The y-axis records the average BMI in each age bucket. The reason we

	Data Set 1 K-Means	Data Set 1 Hierarchical	Data Set 2 GMM	Data Set 2 Hierarchical
2 Clusters	[2506, 60] [2334, 70]	[3693, 100] [1147, 30]	[8652, 2147] [564, 131]	[4487, 1542] [4729, 736]
3 Clusters	[2158, 55] [1855, 49] [827, 26]	[1460, 41] 1147, 30] [2233, 59]	[6560, 1614] [2649, 663] [7, 1]	[4729, 736] [2687, 834] [1800, 708]
4 Clusters	[1566, 32] [1711, 45] [680, 21] [883, 32]	[2233, 59] [987, 28] [473, 13] [1147, 30]	[561, 130] [6003, 1484] [2645, 663] [7, 1]	[2687, 834] [3025, 198] [1800, 708] [1704, 538]
In each cell, each tuple represents the number of non-obese and obese children in a cluster. *More results are available in GitHub*				

Table 2: In-cluster obese/non-obese ratio check

used BMI instead of BMI Percentile was the latter has huge missing data from 0 to 12 months. Also, we did not use BMI z-scores because it was not included in the second data at the time.

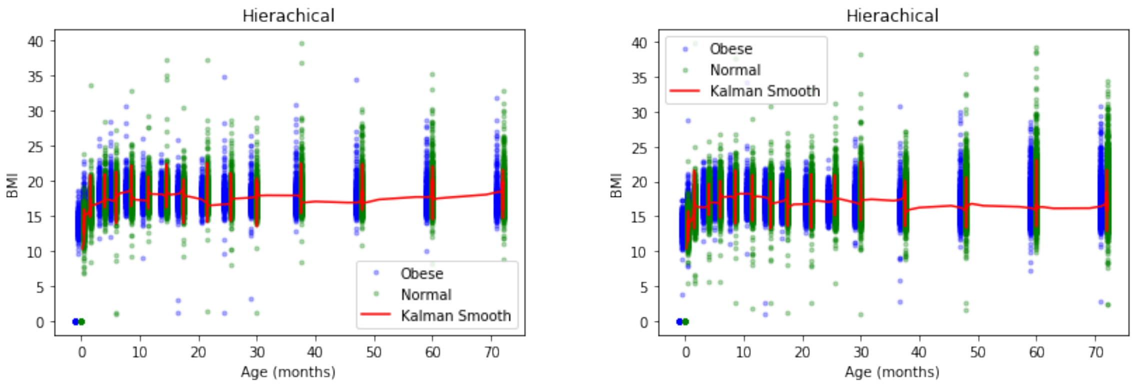


Figure 5: Kalman smoothing and obese/non-obese distribution

We do not see a significant difference in trajectory. Furthermore, the distributions of obese/non-obese at each time point are almost identical, which means the 2-hierarchical clustering did not effectively separate obese from non-obese children.

4.3 Time-Series Based Model

4.3.1 Goal of Time-Series

In theory, traditional time series models may not be a good fit here, since the available time buckets from 0-2 year-old does not seem adequate for predicting the obesity around 5 year-old. Still, in this section we are curious about if we can achieve a predictable BMI curve that extends to 4.5-5.5 year-old by using more buckets (as described in section 4.2) in addition to buckets in 0-2 year-old. For each child, we only have limited number of time stamp data and there were usually some months data missing. It implies time series model cannot produce a convincing fitting curve for each kid. However, fitting all kids' data to get a single prediction line cannot help explain anything. We assumed that if we have clustered data, the results would be more meaningful. Even though clustering task does perform well, in this task, we only take 2 clusters as an example to solve the main task. Based on the previous hypothesis, we clustered data first and fit data in same group to get a prediction line. The x-axis is months, and the y-axis is BMI measurement. The reason why we chose these two features as axes is that we applied time series technique, but BMI itself does not have a direct relationship with time.

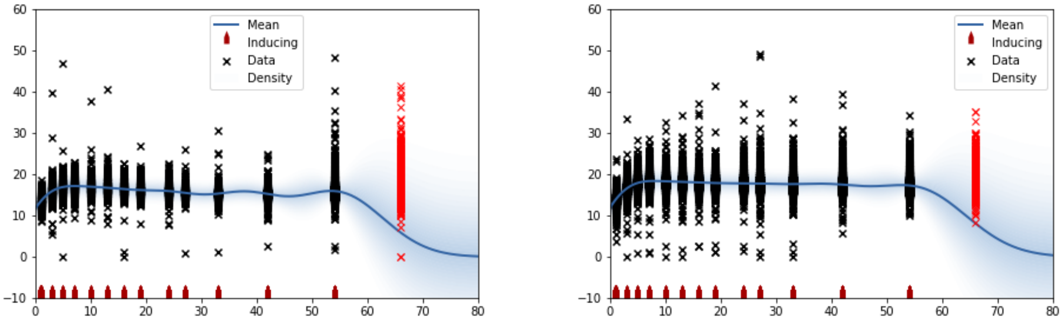
4.3.2 Evaluation

Our model only predicts one datapoint at 5.5-year-old (66 months) for all kids in the same group. But around the curve, the blue shadow area indicates the confidence interval of the prediction line. According to the Central Limit Theorem, kids' BMI are independent, so their values at certain time stamp should move toward a normal distribution. Thus, if the prediction is precise enough, most of data should locate in the blue area. In other words, if the model can really capture childrens growing trajectory, we expect to see more red crosses in the blue shadow.(Red crosses are childrens true BMI on 66 months). But the confidence interval (CI) should be in a reasonable range, otherwise, the curve is not convincing.

4.3.3 Results

As in section 4.2, we chose K-Means and GMM as clustering methods, then tried Gaussian Process with different kernels (RBF kernel, Linear kernel, and RatQuad kernel) to draw growth trajectories. The Python package we have used is `GPy` (GPy since 2012). Gaussian Process is very useful at decision making and forecasting in time series modelling (Roberts et al. 2013). Using different kernels could help capture different covariance functions between variables (Kanagawa et al. 2018). Since data are centered at discrete time stamps, potential relationship between age and BMI is not obvious. We tried different kernel methods to figure out which one is optimal.

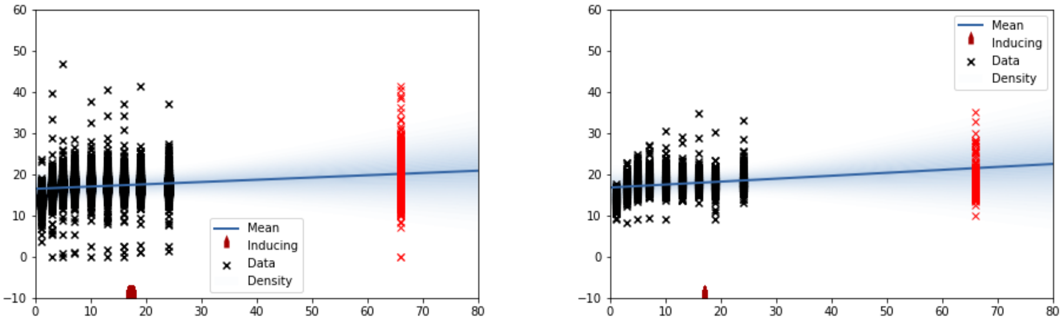
- **RBF Kernel:** RBF kernel, based on Radial basis function, is would like to predict an arc curve, instead of a straight line. So if we only input early months data, the curve would decline early. So for this kernelized function, we had to feed all 0-54 months data to predict. Otherwise, the curve would have large CI to ensure that target BMI would be in the blue shadow. And in this case, K-Means clustering had a better performance than GMM did.



(a) (b)

Figure 6: Best RBF Kernel at 54 Months (2 Groups Clustered by K-Means)

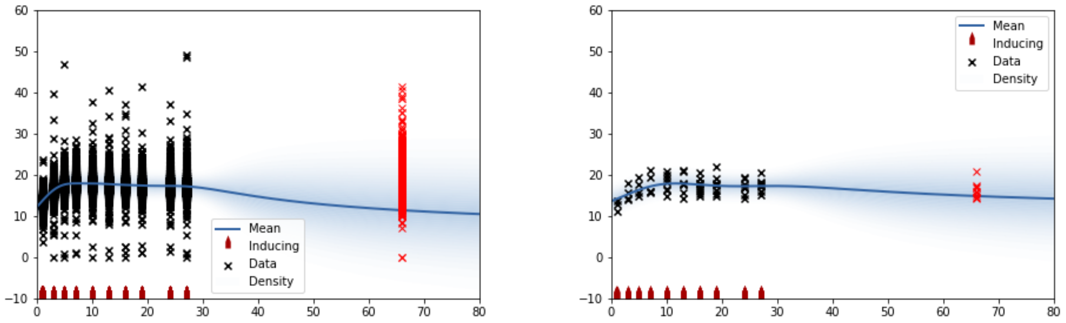
- **Linear Kernel:** It is actually a combination of linear kernel, bias and white noise. The linear kernel regression line crosses the center point in each time stamp. The advantage of it is that it can predict the center of target BMI, and has smaller CI, even if we only fed 0-24 months data. However, if our clustering is not good enough like GMM in this case, and real target BMI data would deviate very much from prediction much.



(a) (b)

Figure 7: Best Linear Kernel at 24 Months (2 Groups Clustered by K-Means)

- **RatQuad Kernel:** Its performance is like combining RBF and linear. Results of feeding 27 months data is acceptable.



(a)

(b)

Figure 8: Best RatQuad Kernel at 27 Months (2 Groups Clustered by GMM)

4.3.4 Analysis of Time Series Performance

- **Clustering Methods:** GMM would like to have more unbalanced clusters than kmeans. If we have more time, we would like to study what does these clusters mean in the future.
- **Regression methods:** If we consider to take an average BMI for all kids in same group, we should take linear kernel. However, if we allow CI of prediction in a reasonable range, RatQuad Kernel is better. RBF kernel did not have good performance in both areas in this task.
- **Model Performance:** For different groups, we hardly see any difference between their trajectory lines. This may be because we have hundreds data points at each time stamp and their latent trajectories may be offset by variance in inputs.

4.4 Neural Network Based Models

To extract more information and simulate the trend of growth in physical characteristics of the children, we use RNN (GRU) based neural network for obesity classification, and RNN encoder-Attention RNN decoder structure for sequence to sequence generation.

4.4.1 Dataset Definition

Based on our original dataset and our handcrafted data, we got 24 time series features. The features are mainly about characteristic data of children. One reason for the selection is those features are more dense thus could provide more stable information stream for our model. For example, a hospital might expect a parent to take the child for examination every month, but most of them only do this irregularly after a child reaches the age of 2. Thus we divide the months into time buckets, e.g. consider 24-27 months as one bucket, and fill in the maximum characteristic data in each feature as representation of this whole time period, thus get one time series feature for this period. Some examples of those features are: ”BMI Percentile”, ”PainScale” and ”Height”.

Our training data includes 9 time slots for each feature, from 0 month to 24 months age, which are: [’avg0to1’, ’avg1to3’, ’avg3to5’, ’avg5to7’, ’avg7to10’, ’avg10to13’, ’avg13to16’, ’avg16to19’, ’avg19to24’]. Each vector in a time slot should be fed in a cell of GRU as input. Since they are float data, we did not perform any embedding for them, and performed normalization in initialization, and batch normalization after each cell in GRU.

Considering we built two models for classification and sequence prediction separately, the target for classification model is the obesity status of a children in the age of 5, with 0 means no obesity, and 1 means obesity; the target for sequence prediction model is another series of time slots, from 24 months to 78 months: [’avg24to27’, ’avg27to33’, ’avg33to42’, ’avg42to54’, ’avg54to66’, ’avg66to78’]. Each vector in a time slot should be a target for the output of corresponding cell in decoder, and used as criteria in loss calculation and back propagation. We also performed normalization with recorded μ and σ for each feature, to recover the output to scale data as results.

4.4.2 Classifier with GRU Encoder for Feature Extraction

The purpose of using a RNN model as classifier is not about improving accuracy or AUC of the classification problem. We want to use the 4.5-5.5 age obesity data as training target to get a good Feature Extractor, and use some hidden states as feature vector for each child, to help find clusters for the children.

With this objective, we tried to build a strong model but train it to a under-fitting condition thus we could get a relatively unbiased feature extractor.
The structure of GRU based classifier is as shown in Figure(9).

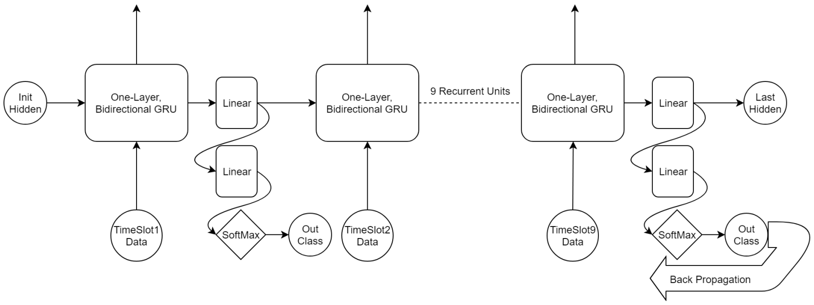


Figure 9: GRU Classifier

We initialize the first hidden state with random floats between 0 and 1. After normalization, feed each time slot vector into GRU, and use two linear layers to extract class information from the last hidden state and extract class probability with SoftMax. The output class will be used in calculating loss between target class and back propagation. We used NLL Loss. The parameters of our NN are: GRU: 24*30, First Linear: 30*30, Second Linear: 30*2.

Although some previous research show its not necessary to use bidirectional GRU in time series problems, we found it performs a little better and kept the structure.

Although the model predicts the result very well, with validation accuracy of 0.8417 and shows theres more information in time series than just flattened data, the Figure(10) shows those extracted last and second last hidden state did not help much in clustering under the current way of evaluation. Further works, including in-cluster obese/non-obese ratio check, corroborated this observation.

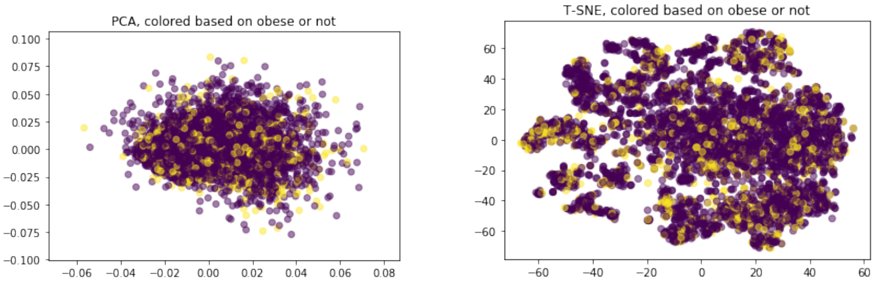


Figure 10: PCA and T-SNE Clustering with Hidden States

4.4.3 GRU and Attention Based Encoder-Decoder Seq2Seq Model

In all previous works, we only aimed to predict an individual target in the future. Here, we also want to build a Seq2Seq model to predict a sequence of future vital characteristics of children. Such prediction, if conducted accurately, could bring new information about ”expected trends” of children. The structure of Seq2Seq model, modified from a neural machine translator (Luong et al. (2015), but with significant changes to make it suitable for float time series problem) is as shown in Figure(11) The encoder part is similar to the GRU Classifier and therefore we omit a repetitive explanation here.

With similar procedure as in GRU classifier, we feed 0-24 months data into the encoder. The output of encoder has two parts: output_data and output_hidden, which are concatenation of all output series data and the last hidden state. To get additional information from the former sequence, we feed both the hidden layer from last cell and output_data into the attention layer to get a 1*9 attention weight, then multiply it with the output_data and concatenate it with input vector in each time slot between 24-78 months, and feed it to the decoder GRU. The next input of decoder GRU

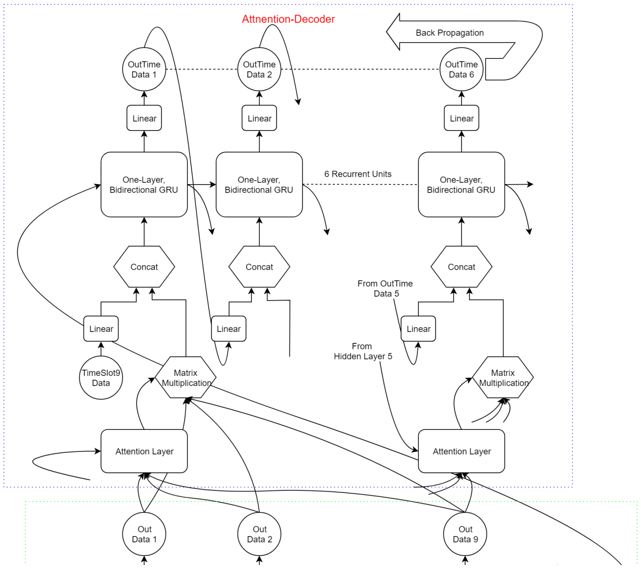


Figure 11: Attention and Decoder in Seq2Seq model

will be concat of last output and weighted output_hidden. We used bidirectional in both encoder and decoder GRU. We compare the result time series with the target time series to get loss and perform back propagation. As target is a vector of floats, we tried both KLDivLoss and MSELoss as criteria. We also used teacher forcing to accelerate convergence.

We tried three kinds of inputs and targets: the whole feature vector input to generate whole feature vector target, whole input to generate height and weight, and H & W to generate H & W, so we could obtain BMI based on $BMI = W/H^2$. We also did a lot of hyperparameter tuning. However, the comparation of training curve under KLDivLoss and MSELoss is more interesting as in Figure(12).

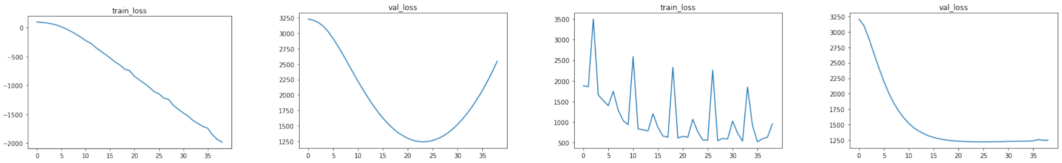


Figure 12: Training Curves Seq2Seq under KLD Loss(left) and MSE Loss(right)

In training, we use only MSE Loss as score function for validation set, but both KLD and MSE Loss as cost function for training. It seems that KLD Loss could make the model highly overfit on training, but validation loss starts to raise soon. On the other hand, MSE Loss training could somehow prevent overfit on this kind of float-fitting problem.

We trained the data with 5 epoches and learning rate 2e-6. The best model come from target set only with weight and height vectors with MSE Loss. When predicting all 24 features as targets, although the loss seems smaller in value, but thats because the target is more sparse in dimension, and most of them are zero. The parameters of our best model are:

- Encoder:
  Input Linear: 24*200, GRU: 200*200
- Attention:
  Input Layer:[1*200+9*200]*200, Weight Calculator: 200*9, Output: 9.
- Decoder:
  Input Linear: 24*200, GRU: 400*200, Output Linear: 200*24.

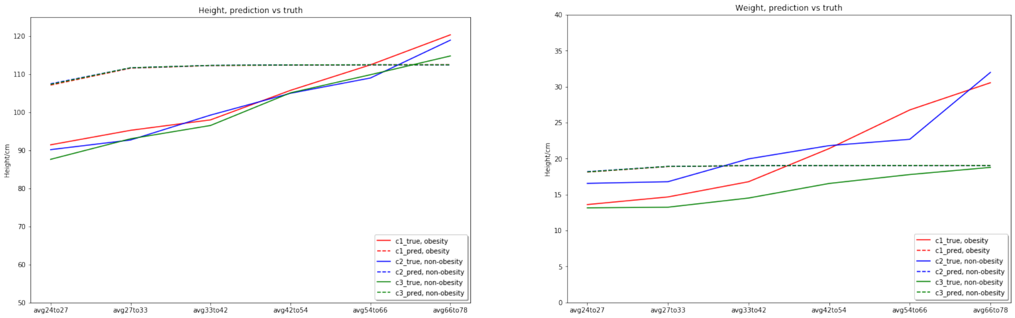


Figure 13: Prediction v.s. Truth of Height and Weight, Three Examples

Figure(13) shows three instances each in prediction v.s. truth. Both models could not capture the information of growth very well, probably because of the dataset is too small (only 4000 instances), and most of the features are very sparse with huge amount of zeros. However, the model trained with only weight and height information as targets could predict a more credible scale for the characteristics of children. Overall, the seq2seq model trained should not be considered as a good predictor of the childrens growth curve of height and weight, we need more information and datapoints to get a strong enough model to obtain different patterns in growth of children.

5. Discussions

5.1 Conclusions

Generally speaking, our project can be divided into two parts: clustering and trajectory fitting (unsupervised), and trend-related modeling (supervised). For the first part, the main observation we acquired is that the ”natural categories and the corresponding trajectories” do not exist at all, at least under the current evaluation criteria. Still, it may be possible that, based on other features besides obesity, we can separate children into categories, and, with the help of more advanced clustering algorithms, we can get different growth trajectories among all groups. Also, more exploration on the hidden state from GRU model may bring new insights to the clustering problem, although our current attempts did not present fruitful results.

For our second part, first, although we have discussed why traditional time series models do not fit well here, our attempts of using more buckets could be an introduction to a new topic in this study. Indeed, our current attempt is yet a ”prediction problem based on strict definition”, and more detailed problem formulation may be required to continue. Neural Network based model produced good result on the original problem set, but on the ”trend prediction” problem, obviously it needs refinement to achieve better results.

Two things, as the bottlenecks for our work, worth mentioning here are the volume and sparsity of data. As discussed, the small volume of data potentially hinders the performances of both time series model and GRU model, for they both require a relatively large volume of data. Also, sparsity, sometimes equivalent to small data volume, brought challenges to modeling as well. Had we acquire more data for this study, we were confident that a better result in the second part can be achieved.

5.2 Future Work

Several proposals could be taken as future references. First, as we discussed above, we may have to change the ”expectation” of our clustering section. It is likely that the hope to separate obese and non-obese children is not reasonable for the given data set. On the contrary, the ”true natural clusters” are possibly categorized by other features, such as disease or demographic indices. Unfortunately, considering the huge amount of features, even those features not included in our data set, such ”guess and try” work can be tedious.

Also, a plausible future improvement can be done with our Seq2Seq model. Despite the unconvincing performance at this point, we are optimistic about the space for future refinements, either by feeding more data or by reconstructing some components of it. After all, we have already had a considerably clear problem formulation and evaluation criteria for this model.

6. Miscellany

Relevant codes are available at
https://github.com/NYU-CDS-Capstone-Project/Obesity_Prediction

Appendix

Kalman smoothing

The Kalman smoothing can be applied on the linear Gaussian Hidden Markov models. Smoothing works as it computes the posterior distribution of the latent state provided that the full sequence of observations are given. This satisfies our experimental setting for the project. That is, we would like to draw the trajectory (latent states) of children given full BMI Percentile observations (Savin 2018).

The Python package we have used in this project is `pykalman` (Duckworth 2012).

References

Julian Alston and Abigail M. Okrent. *The effects of farm and food policy on obesity in the United States.* Palgrave Macmillan, 2017.

Daniel Duckworth. pykalman. `https://github.com/pykalman/pykalman`, 2012.

GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

Joseph F Hagan, Judith S Shaw, and Paula M Duncan. *Bright futures: Guidelines for health supervision of infants, children, and adolescents.* Am Acad Pediatrics, 2007.

Robert Hammond, Rodoniki Athanasiadou, Silvia Curado, Yindalon Aphinyanaphongs, Courtney Abrams, Mary Jo Messito, Rachel Gross, Michelle Katzow, Melanie Jay, Melanie Jay, and Brian Elbel. Predicting childhood obesity using electronic health records and publicly available data. Unpublished work, under review., 2018.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf`.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Microsoft. lightgbm. `https://github.com/Microsoft/LightGBM/tree/master/python-package`, 2017.

S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for timeseries modelling. *Philosophical Transactions of the Royal Society (Part A*, 2013.

Christina Savin. Kalman filtering and smoothing derivations. `https://github.com/charlieblue17/pTSAFall2018/blob/master/lectures/kalman.pdf`, 2018.

Asheley Cockrell Skinner, Sophie N Ravanbakht, Joseph A Skelton, Eliana M Perrin, and Sarah C Armstrong. Prevalence of obesity and severe obesity in us children, 1999–2016. *Pediatrics*, page e20173459, 2018.