HANDLING DATA GAPS IN TIME SERIES

USING IMPUTATION

PRESENTED BY

CLARE
JEON

ALF
WHITEHEAD

Klick Health

HELLO!

CLARE
JEON

ALF
WHITEHEAD

Klick Health

CONTENTS

1	What is time-series dataset & how do we use the data?
2	Missing values and imputation - What can we do about that missing data?
3	*Real world example:* Predicting blood glucose level of Type 1 Diabetes patients at 30 minutes horizon

CONTENTS

1 What is time-series dataset & how do we use the data?

2 Missing values and imputation - What can we do about that missing data?

3 *Real world example:*
Predicting blood glucose level of Type 1 Diabetes patients at 30 minutes horizon

Klick Health

TIME SERIES DATA

bitcoin_price_1week_Test - bitcoin_price_Aug2017_1week.csv (605 B) 7 of 7 columns Views

	Date	# Open	# High	# Low	# Close	# Volume	# Market...
1	Aug 01, 2017	2871.3	2921.35	2685.61	2718.26	1,324,670,000	47,321,800,000
2	Aug 02, 2017	2727.13	2762.53	2668.59	2710.67	1,094,950,000	44,950,800,000
3	Aug 03, 2017	2709.56	2813.31	2685.14	2804.73	804,797,000	44,666,400,000
4	Aug 04, 2017	2806.93	2899.33	2743.72	2895.89	1,002,120,000	46,276,200,000
5	Aug 05, 2017	2897.63	3290.01	2874.83	3252.91	1,945,700,000	47,778,200,000
6	Aug 06, 2017	3257.61	3293.29	3155.6	3213.94	1,105,030,000	53,720,900,000
7	Aug 07, 2017	3212.78	3397.68	3180.89	3378.94	1,482,280,000	52,987,300,000

Time series data is:

- A sequence of values of variables
- Measures the same thing over time and stores them in time order

https://www.kaggle.com/team-ai/bitcoin-price-prediction/version/1

TIME SERIES ARE EVERYWHERE IN SIGNAL PROCESSING

$y = f(t)$

f(t) is implicitly defined everywhere, but the data is only sampled.

A common error is to use time series to track discrete events (e.g. events on a purchasing journey). This isn't helpful because there is no "value" between the samples - use a sequence representation instead.

FORECASTING IN HEALTHCARE: GLUCOSE IN TYPE 1 DIABETES

Our study system:
Glucose (Pathogenesis of Diabetes Mellitus)

Daily Change of Glucose Level

DOI: 10.1109/TBME.2015.2512276

Klick Health

FORECASTING IN HEALTHCARE: GLUCOSE IN TYPE 1 DIABETES

DOI: 10.1109/TBME.2015.2512276

Klick Health

TYPE 1 DIABETES

Insulin therapy is the only way for patients.

Accurate blood glucose level prediction can help to figure out when insulin should be injected.

Klick Health

FORECASTING IN HEALTHCARE: GLUCOSE IN TYPE 1 DIABETES

Major Complications of Diabetes

Microvascular

Eye
High blood glucose and high blood pressure can damage eye blood vessels, causing retinopathy, cataracts and glaucoma

Kidney
High blood pressure damages small blood vessels and excess blood glucose overworks the kidneys, resulting in nephropathy.

Neuropathy
Hyperglycemia damages nerves in the peripheral nervous system. This may result in pain and/or numbness. Feet wounds may go undetected, get infected and lead to gangrene.

Macrovascular

Brain
Increased risk of stroke and cerebrovascular disease, including transient ischemic attack, cognitive impairment, etc.

Heart
High blood pressure and insulin resistance increase risk of coronary heart disease

Extremities
Peripheral vascular disease results from narrowing of blood vessels increasing the risk for reduced or lack of blood flow in legs. Feet wounds are likely to heal slowly contributing to gangrene and other complications.

DOI: 10.2210/rcsb_pdb/GH/DM/monitoring/complications

Klick Health

MANAGE TYPE 1 DIABETES

CGM = Continuous Glucose Monitoring

Photo courtesy: U.S. Food and Drug Administration

Klick Health

REAL WORLD MISSING DATA IN TYPE 1 DIABETES PATIENTS

Continuous Glucose Monitoring device is not perfect!

Faulty sensors, incomplete records, mistakes in data collection, unavailability to report certain information (at a given time point) and others...

Klick Health

CONTENTS

1 What is time-series dataset & how do we use the data?

2 Missing values and imputation - What can we do about that missing data?

3 *Real world example:*
Predicting blood glucose level of Type 1 Diabetes patients at 30 minutes horizon

Klick Health

THREE QUESTIONS

1	Which imputation methods are available?
2	How do we use them?
3	How do we evaluate imputation methods?

Q1. IMPUTATION METHODS FOR TIME SERIES DATA

UNIVARIATE TIME SERIES IMPUTATION	MULTIVARIATE TIME SERIES IMPUTATION
Mean (Median)	K-Nearest Neighbors
Last Observation Carried Forward	Random Forest
Linear Interpolation	Multiple Singular Spectral Analysis
Polynomial Interpolation	Expectation-Maximization
Kalman Filter	Multiple Imputation with Chained Equations
Moving Average	
Random	

Q1. IMPUTATION METHODS FOR TIME SERIES DATA

Canonical ML/DL modeling
Use both past and future values

Real-time prediction (Online mode)
Use only past value

Linear Interpolation Polynomial Interpolation Kalman Smoothing Moving Average K-Nearest Neighbors Random Forest Multiple Singular Spectral Analysis Expectation Maximisation Multiple Imputation with Chained Equations	Last Observed Carried Forward Mean (Median) Random Empirical Based

Q2. HOW DO WE USE IMPUTATION METHODS

Today's Sample Data: 24 hours Blood glucose levels

Klick Health

SETUP

PYTHON

```
import pandas as pd
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt

from fancyimpute import knn
import rpy2.robjects.packages import importr
imputeTS = importr("imputeTS")
kalman_StructTs = robjects.r['na.kalman']
kalman_auto_arima = robjects.r['na.kalman']

df = pd.read_excel('glucose.xlsx')
df.set_index('Minute_of_Day', inplace=True)
missing_minutes = \ list(df[df['Inferred_Glucose'].isna()].index)
```

R

```
library (xlsx)
library (imputeTS)
library (pracm)
library(spatialEco)
library (VIM)

df <- read.xlsx("glucose.xlsx", sheetName = "Sheet1")

df[1] <- NULL
rownames(df) <- df$Minute_of_Day

missing_minutes <- df[which(is.na(df$Inferred_Glucose)),
'Minute_of_Day']
```

Klick Health

“TRADITIONAL METHOD”

Remove entire missing values.

When there are small fraction of missing values.

This is valid if the data is missing completely at random.

PYTHON

```
df[...].dropna()
```

R

```
df <- na.omit(df)
```

LAST OBSERVATION CARRIED FORWARD

Last observed value of a variable is used for all subsequent.

Can be either conservative or liberal depending on the context.

PYTHON

df[...].fillna(method = 'ffill')

R

imputeTS::na_locf(x)(x, option = ‘locf’)

MEAN VALUE

On average, the signal's value is the mean.

PYTHON

df[...].fillna(df[...].mean())

R

imputeTS::na_mean(x, option = "mean")

LINEAR INTERPOLATION

What the graphing libraries do by default - draw a line connecting the two points at either end.

Can you do this while data is being collected? Nope.

PYTHON

df[...].interpolate(method='linear')

R

imputeTS::na_interpolation(x, option = “linear”)

NEAREST NEIGHBOR

We can use the measured point nearest to the missing sample as the estimate.

PYTHON

```
df[...].interpolate(method='nearest')
```

R

```
pracma::interpl(x, y, method = 'nearest')
```

POLYNOMIAL INTERPOLATION

What if we take into account the trajectory of the signal at loss and reacquisition? We could fit a curve using the derivatives, which would be a polynomial.

Stineman is a special type of polynomial fit.

PYTHON

df[...].interpolate(method='polynomial', order=2)

df[...].interpolate(method='polynomial', order=3)

R

spatialEco::poly.regression(df$Inferred_Glucose, s = 0.2, impute = TRUE, na.only = TRUE)

SPLINE INTERPOLATION

Splines are an alternate way to fit “curvy lines” to data. They assume the use of “knots,” or fixed points, around which a semi-rigid curve is bent. This comes from pencil-and-paper drafting techniques.

PYTHON

```
df[...].interpolate(method='spline', order=2)

df[...].interpolate(method='spline', order=3)
```

R

```
imputeTS::na_interpolation(x, option = “spline”)
```

MOVING AVERAGE IMPUTATION

Missing values are replaced by moving average values. The mean in this implementation taken from an equal number of observations on either side of a central value.

PYTHON

```
df[...].rolling(4).mean()
```

R

```
imputeTS::na_ma(df[...], k = 4)
```

KALMAN SMOOTHING IMPUTATION

Filling missing values by estimating a joint probability distribution over the variables for each timeframe.

PYTHON

kalman_StructTs(df[...], model = "StructTS")

kalman_StructTs(df[...], model = "auto.arima")

R

imputeTS::na_kalman(df[...],model = 'StructTS')

imputeTS::na_kalman(df[...],model = 'auto.arima')

Klick Health

KNN IMPUTATION

For k-Nearest Neighbor imputation, the missing values are based on a kNN algorithm. In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate.

PYTHON

```
KNN(k = 5).fit_transform(df)[:, 0]
```

R

```
VIM::kNN(df[...],
variable = 'Inferred_Glucose', k = 5)
```

EMPIRICAL BASED IMPUTATION

Assumption: Patients have regular life schedule

Klick Health

Q3. HOW DO WE EVALUATE IMPUTATION METHODS

1	Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N} \lvert y_i - \hat{y} \rvert$
2	Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
3	Pearson's Correlation Coefficient (R^2)	$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$ Where, $\hat{y}$ – *predicted value of y*; $\bar{y}$ – *mean value of y*
4	**Rank Product**	$RP(g) = \sqrt[k]{\left(\prod_{i=1}^{k} r_{g,i}\right)}$

Geometric mean of the ranks of imputation methods determined by MAE, RMSE and R^2

Generalize performance evaluation

Klick Health

CONTENTS

1 What is time-series dataset & how do we use the data?

2 Missing values and imputation - What can we do about that missing data?

3 *Real world example:*
Predicting blood glucose level of Type 1 Diabetes patients at 30 minutes horizon

PREDICTING BLOOD GLUCOSE LEVELS OF DIABETES PATIENTS EXPERIMENTS IN DATA IMPUTATION

Our team were selected as finalists and we presented our work at IJCAI 2018 in Stockholm.

Predicting Glycemia in Type 1 Diabetes Patients: Experiments with XGBoost

Cooper Midroni, Peter J. Leimbigler, Gaurav Baruah, Maheedhar Kolla, Alfred J. Whitehead, Yan Fossat
Klick Inc., 175 Bloor Street East, Toronto, Ontario, Canada
{cmidroni,pleimbigler,gbaruah,mkolla,awhitehead,yfossat}@klick.com

Abstract

Type 1 diabetes patients must self-administer insulin through injections or insulin-pump therapy, requiring careful lifestyle management around meals and physical activity. Accurate blood glucose prediction could increase patient quality of life, and foreknowledge of hypoglycemia or hyperglycemia could mitigate risks and save lives. For the 2018 BGLP Challenge, we experiment primarily with XGBoost to predict blood glucose levels at a 30-minute horizon in the OhioT1DM dataset. Our experiments show that XGBoost can be a competitive predictor of blood glucose levels, as compared to prior research, and that feature signals from different sources contribute in varying capacity for improved predictive ability of XGBoost.

1 Introduction

Diabetes affects over 400 million people worldwide [World Health Organization, 2016], with near 5% of diabetics suffering from type 1 diabetes (T1D) [American Diabetes Association, 2018]. Patients with T1D are incapable of producing insulin, a hormone generated by the pancreas, which acts as the primary regulator of blood glucose metabolism. This dysfunction can lead to both hypoglycemia (low blood sugar) and hyperglycemia (high blood sugar), resulting in a significant patient burden to regulate carbohydrate consumption and supplemental insulin delivery. Hyperglycemia can lead to medical complications such as vision loss and kidney failure, and increases risk of heart disease and stroke. Hypoglycemia can lead to loss of consciousness and even death.

An increasing number of T1D patients are adopting insulin pump therapy, wherein a wearable device releases insulin subcutaneously to mimic pancreatic response. Current insulin pumps require patient input on carbohydrate intake and approval of each recommended insulin dose. Driven by the outstanding need for closed-loop insulin therapy, the notion of an artificial pancreas has gained traction in diabetes-related research [Graf *et al.*, 2017; Juvenile Diabetes Research Foundation, 2018].

The dysregulation of blood glucose in T1D patients is further complicated by daily variations in the magnitude and timing of meals, physical activity, and insulin self-administration. This, along with the altered pharmacokinetics of subcutaneous insulin, add further layers of complexity to the task of predicting blood glucose. As such, the Blood Glucose Level Prediction Challenge represents an important step toward the realization of an artificial pancreas. Herein lies the objective of restoring homeostasis through accurately dosed insulin, and the creation of a model which captures the complexities of the disease. This challenge was particularly motivating for us, not simply from the perspective of predictive modeling, but also for the potential applications in providing tangible benefits to T1D patients.

To predict glucose at a 30-minute time horizon, we processed the raw features of the OhioT1DM dataset [Marling and Bunescu, 2018] to create 3 different feature sets, and experimented with gradient-boosted trees [Chen and Guestrin, 2016a] (XGBoost), random forests [Breiman, 2001], and recurrent neural network variants. We find that:

- XGBoost performs on par with prior models [Bunescu *et al.*, 2013; Mirshekarian *et al.*, 2017] for this task.
- Many of the provided features do not contribute to improved predictive performance. In essence, when using XGBoost, past glucose is the most important predictor of future glucose.
- Using ϵ-insensitive loss for training LSTMs improves predictive performance compared to mean squared error.

2 Data

The OhioT1DM dataset comprises 19 features collected from 6 patients with T1D [Marling and Bunescu, 2018]. Patients wore Medtronic 530G insulin pumps, Medtronic Enlite continuous glucose monitors (CGM), and Basis Peak fitness wristbands. 8 weeks of data were provided per patient, of which the final 10 days were provided separately as a test set.

2.1 Data Analysis

Due to the heterogeneity of the data, we grouped features according to their frequency:

- *one-off data*: intermittent measurements with no fixed sampling frequency or duration. (e.g., finger stick glucose, insulin bolus time and dose, sleep times and quality, work intensity, exercise intensity and duration, meal

Klick Health

BUILDING PREDICTIVE MODELS FOR BLOOD GLUCOSE LEVEL

Klick Health

BUILDING PREDICTIVE MODELS FOR BLOOD GLUCOSE LEVEL

Klick Health

COMPARISON OF METHODS FOR MISSING DATA IMPUTATION

Klick Health

THE EFFECT OF MISSING DATA IMPUTATION ON BLOOD GLUCOSE LEVEL PREDICTION

Klick Health

ENSEMBLE MODELING

Klick Health

ENSEMBLE MODELING

Klick Health

FOR YOUR WORK

1. We've shown 11 approaches for missing data that you can chose from

2. Important questions to ask:
   Is your dataset univariate or multivariate?
   Do you need the ability to make real-time predictions with missing data?

3. Pick error metric(s) and test the appropriate imputation methods using those metrics

4. For best results, ensemble the best-performing methods

KLICK – WHO WE ARE

FAST COMPANY'S BEST WORKPLACES FOR INNOVATORS 2019 HONOREE

Klick Health

THANK YOU.

https://github.com/KlickInc/datasci-strata-talk-missing-data

PRESENTED BY

CLARE
JEON

ALF
WHITEHEAD

Klick Health

Rate today’s session

Session page on conference website

O’Reilly Events App