**CMPE541 - PROJECT #1**

The objective of this project is:

* Work with a real-life dataset
* Practice Data Transformation
* Practice Noise removal and data summarization techniques in conjunction
* Provide and Implement a data-specific tool for the given real-life dataset
* Have a rough understanding of data cube approach (ahead of time in our lectures)

In the below link, you can find a real life diabetes dataset.

https://www.dropbox.com/sh/k9rk1rdpnn7b84g/AACD7zyOnMPXN6J-7H8U8v5ba?dl=0

alternatively you can use the following link in order to download it:

https://archive.ics.uci.edu/ml/datasets/diabetes

The dataset contains an extensive information on 70 diabetes patients in multiple files. For each patient, a catalogued set of attributes / events and their associated time is given at each line. As the final motivation of the project, we aim to find whether we can detect blood glucose information by relating it to some (thing?) of these attributes / events.

In this assignment, first the dataset will be described, then a short information about diabetes is provided. Next, some ideas about preprocessing the dataset is given as suggestions and the assignment concludes.

The document collection consists of 70 files, each belonging to a different patient. Each line in the files represent a measurement and each line consists of four values:

(1) Date in MM-DD-YYYY format

(2) Time in XX:YY format

(3) Code

(4) Value

The Code field is deciphered as follows:

33 = Regular insulin dose

34 = NPH insulin dose

35 = UltraLente insulin dose

These are insulin medication taken by the patients, categorized as the onset and effective time of the medication.

48 = Unspecified blood glucose measurement

57 = Unspecified blood glucose measurement

58 = Pre-breakfast blood glucose measurement

59 = Post-breakfast blood glucose measurement

60 = Pre-lunch blood glucose measurement

61 = Post-lunch blood glucose measurement

62 = Pre-supper blood glucose measurement

63 = Post-supper blood glucose measurement

64 = Pre-snack blood glucose measurement

These are blood sugar level measurements. Usually code 59 is considered the most standardized measurement for blood sugar (where ranges in 70-100 is considered normal), however the blood sugar measurements can also be taken 2 hours after a meal and is also considered standardized (100 - 125 is considered normal). Unspecified measurements correspond to any-time measurements that can show varied response, and is usually considered not reliable.

65 = Hypoglycemic symptoms

Too low blood sugar is considered hypoglycemia and is considered very dangerous. This condition is further catalogued in the data with a code.

66 = Typical meal ingestion

67 = More-than-usual meal ingestion

68 = Less-than-usual meal ingestion

These fields show categorized food intake.

69 = Typical exercise activity

70 = More-than-usual exercise activity

71 = Less-than-usual exercise activity

These fields show categorized organized exercise activity.

72 = Unspecified special event

Any other event. Not specified. We don't know what it is, should be removed.

**DIABETES SHORT**

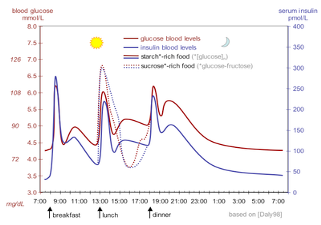
Domain knowledge is important, sometimes even more important than programming and data mining knowledge. So let's read a little on diabetes:

Diabetes is a progressive and life-long disease due to the decline of the insulin hormone secreted by pancreas or deficiency of the utilization of the insulin hormone. The effects of diabetes manifests itself as a reason of cells being unable to process the glucose in the bloodstream. As a deficiency having a close relationship with the glucose levels in the bloodstream, the diabetes usually increases overall level of blood sugar in the body over long periods of time, and thus causing many additional illnesses. Additionally, the low blood sugar levels can also distort bodily functions of human beings. To this day, there is no known cure for diabetes and thus, is a disease that requires constant monitoring throughout a patient's life.

It is a well established fact that several personal traits such as age, height, weight and genetic disposition of a person effect the blood glucose levels. In this study, such information is kept anonymous and hence, unavailable (we have to make due)

What we are really interested is the fact that blood glucose level is actually a trend; if it is high that person's trend is high, and if it is low that person's trend is low. Similarly, the food consumption and sports activities (a.k.a. being healthy) also correlate with blood sugar levels.

Finally, diabetes is a metabolic activity. Below, the figure shows a normal blood sugar level projection for a person. Notice how it changes: it peaks during mid-day and goes low at nighttime and early daytime.



Finally, it is also a known fact that blood sugar levels and projection throughout the day shows personalized markers. That is, each person has a specific blood sugar level graph (although similar to the one above.)

The aim of this project is to promote intelligent, conscious, and effective self-monitoring by means of providing patients intelligent mechanisms and models to predict their blood glucose levels at any time of day through machine learning. This way, it is expected that the patients will have an educated idea as to if and when there is an unexpected behavior in their blood sugar levels and can seek professional aid before any permanent harm is done on their health.

**How to Do Blood Sugar Level Prediction - Preliminaries**

First off, we cannot work with 70 different files. We need to consolidate them. Also we may need additional markers associating different measurements for food consumption, blood sugar measurements and activity. Below are the suggestions I am making. You can use any programming language, even use a professional database tool to do it; You may also deviate from my suggestions for structuring the data. The aim is having correct predictions at the end of project #3, so it is up to you. Here are some of my suggestions:

We need to consolidate different patient files into a single file/database structure so that we don't need to deal with multiple files all the time. Additionally, we need to differentiate users in this file with some marker. Last, the time structure is not a usable one, thus convert times into a numeric form.

What we can do is a 4-file structure. The files and their contents are as follows:

* **.info:** Contains two numbers; how many users there are and how many total attributes are recorded in the other files.
* **.users:** this file has as many lines as users in our dataset. The line number associates user with a user ID (patID - patient ID). In each line there is a single number representing how many entries that user has. (This file can be extended in order to keep how many blood measurements, food intakes and exercise activities that person has and etc.)
* **.attrs:** One line is associated with each attribute. at each line there are 5 numeric entries: attribute code, [0,1] is it a blood sugar measurement, [0,1] is it a standard measurement, [0,1] is it a meal measurement, [0,1] is it an exercise measurement.
* **.dat:** at each line there is a time stamp, a patient ID, an attribute code, and attribute value. The total number of lines stored in the .info file.

Last, time in the original data is not suitable for comparison. We might like to convert it into linux timestamp (google it !).

As a last task, sort the .dat file with respect to time.

More information on how we create features for our machine learning models in Project #2.

Note: You need to write a 2-4 page report on