Capstone Proposal

Machine Learning Engineer Nanodegree

Derek Cheng
April 1st, 2018

Domain Background

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Over time, having too much glucose in your blood can cause health problems, such as heart disease, nerve damage, eye problems, and kidney disease. Diabetes is caused mainly due to the consumption of highly processed food, bad consumption habits, and so on. According to WHO, almost half of all deaths attributable to high blood glucose occur before the age of 70 years, and diabetes will be the seventh leading cause of death in 2030.

Since diabetes is also an inherited disease in my family, I would like to know whether machine learning techniques can be utilized to predict the occurrence of diabetes. By building a prediction model for diabetes, we may have further understanding on which factors are mostly related to this kind of disease, and therefore find a way to prevent or delay the onset of diabetes beforehand.

Problem Statement

The goal is to use an off-the-shelf diabetes dataset to establish a binary classification model that can correctly predict the whether a person has diabetes or not. The model accuracy should at least exceed the benchmark model and be able to answer the questions such as “which factors are mostly related to the onset of diabetes” and “what the probability of getting diabetes is if the predictor values are given?”

Datasets and Inputs

The “Pima Indians Diabetes Database” dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, and provided by the UCI Machine Learning Repository (can also be downloaded at kaggle). The datasets consist of several medical predictor variables including:

- **Pregnancies**: Number of times pregnant – It is reported that "gestational diabetes mellitus" which is a form of diabetes that occurs only during pregnancy, would let some women continue to have high blood glucose levels after delivery[1].
- **Glucose**: Plasma glucose concentration – diabetes is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period, and therefore glucose would be the most direct indicator of diabetes.
- **BloodPressure**: Diastolic blood pressure (mm Hg) - Having diabetes increases the risk of developing high blood pressure and other cardiovascular problems, because diabetes predisposing the arteries to atherosclerosis[2].
- **SkinThickness**: Triceps skin fold thickness (mm) - Skin thickness is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus (IDDM)[3].
- **Insulin**: 2-Hour serum insulin (mu U/ml) - Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced[4].
- **BMI**: Body mass index (weight in kg/(height in m)^2) - Increased BMI was associated with increased prevalence of diabetes mellitus[5]
- **DiabetesPedigreeFunction**: Diabetes pedigree function - a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject. It utilizes information from a person's family history to predict how diabetes will affect that individual[6].
- **Age**: Age (years) – Though the age at which someone develops the condition depends on many differing factors, it has been reported that age greatly increases the chances of developing type II diabetes.
- **Outcome**: Class variable (0 or 1)

Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. There are 768 persons contained in this dataset, 500 are labeled as 0 (non-diabetic) and 268 as 1 (diabetic), so the two classes are a little imbalanced. Since the number of dataset is limited, the whole data of the two classes will be utilized even though they are imbalanced, and the whole dataset will be split into training and testing set with a test size of 0.2. During the model training, 10-fold cross validation will be used to efficiently use the limited data in training set and provide a more accurate estimate of out-of-sample accuracy. After the training and hyper-parameter tuning, the previously separated testing set will be used to evaluate the model performance by calculating the evaluation metrics.

[1] https://www.diabetesaustralia.com.au/gestational-diabetes

[2] https://www.webmd.boots.com/diabetes/guide/diabetes-bp

[3] Collier, A., Patrick, A. W., Bell, D., Matthews, D. M., Maclntyre, C. C., Ewing, D. J., & Clarke, B. F. (1989). Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients. Diabetes Care, 12(5), 309-312.

[4] Gardner, D. G., Shoback, D., & Greenspan, F. S. (2007). Greenspan's basic & clinical endocrinology. McGraw-Hill Medical,.

[5] Bays, H. E., Chapman, R. H., & Grandy, S. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. International journal of clinical practice, 61(5), 737-747.

[6] Shanker, M., Hu, M. Y., & Hung, M. S. (2000). Estimating probabilities of diabetes mellitus using neural networks. SAR and QSAR in Environmental Research, 11(2), 133-147.

Solution Statement

The solution for this project is to use the diagnostic measures contained in the dataset “Pima Indians Diabetes Database” as inputs to establish a binary classification model that can correctly predict the target variable **outcome** (Boolean variable that represents whether a patient has diabetes) that indicates whether a Pima Indian Woman has diabetes.

Benchmark Model

Logistic regression is the model of choice in many medical data classification tasks, especially for binary classification problem. The power of the study in case of the logistic regression is between 0.80 to 0.857. The logistic regression model calculates the class membership probability for one of the two categories in the data set:

$$P(1|x,\alpha) = \frac{1}{1+e^{-(\alpha\cdot x)}}, \qquad P(0|x,\alpha) = 1 - P(1|x,\alpha)$$

Here $P(1|x,\alpha)$ is the dependence of the posterior distribution on the parameters α. The hyperplane of all points α satisfying the equation $\alpha \cdot x = 0$.

To build a simple logistic model, it is often common to build a correlation matrix first to see which features are highly correlated with the target variable. A simplest choice is to the use the first two features (in this dataset, these are **Glucose** and **BMI**) who have the highest correlation. A simple logistic regression model using these two features should be able achieve a prediction accuracy about 80%[8]

[7] Agresti, A. (2013). Categorical data analysis. John Wiley & Sons.

[8] https://www.kaggle.com/mshirlaw/pima-indians-diabetes-simple-logistic-regression

Evaluation Metrics

Because our dataset shows a little imbalanced between two classes. It would be unsuitable to use the prediction accuracy for quantifying the performance of the benchmark model and the solution model. On the contrary, precision and recall can provide a more accurate measure of success of prediction when the classes are imbalanced, which can be the case in out dataset. Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. The definition of precision and recall are calculated as follows:

True Positive (TP): number of records that are correctly classified.

True Negative (TN): number of valid records that are correctly classified

False Negative (FN): the records are incorrectly classified.

False Positive (FP): the records are incorrectly classified as positive

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Project Design

The tasks involved in the workflow of this project can be summarized as the following:

1. Download the “Pima Indians Diabetes Database” dataset and perform data exploration to gain basic understanding about the potential features.
2. Perform data cleaning to handle invalid data values. To be specific, we need to check whether there are:
   - ✧ Duplicate or irrelevant observations.
   - ✧ Bad labeling of data, same category occurring multiple times
   - ✧ Missing or null data points.
   - ✧ Unexpected outliers
3. Do some basic feature engineering such as one-hot encoding feature transformation for categorical variables and feature scaling if the scales of the features are inconsistent.

4. Train classifiers based on several machine learning algorithms and perform model selection to choose the best model that performs better with the diabetes data set with default parameters. The machine learning algorithms considered at this phase are **Logistic Regression**, **K-Nearest Neighbors**, **Support Vector Classifier**, **Gaussian Naive Bayes**, **Random Forest** and **Gradient Boost**.

5. After selecting the best classification model, do feature selection to select which features most importantly affect the performance of the model and get rid of the features that do not improve the model. There are several selection strategies that would be useful:
   - ✧ **Univariate Feature Selection**: Statistical tests can be used to select those features that have the strongest relationship with the output variable.
   - ✧ **Recursive Feature Elimination (RFE):** RFE works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.
   - ✧ **Principal Component Analysis (PCA):** PCA linearly transform the dataset into a compressed form. We can choose the number of dimensions or principal component in the transformed result.
   - ✧ **Feature Importance:** Bagged decision trees like Random Forest can be used to estimate the importance of features.

6. Model hyper-parameter tuning: perform an exhaustive search like the grid search technique to further increase the model prediction accuracy and other performance metrics.