CS532 Pattern Recognition

Diabetes Data Analysis and Classification

Fan, Zeng (zfan24@wisc.edu)
Mei, Chaoqun (cmei3@wisc.edu)
Zhang, Guiming (gzhang45@wisc.edu)

1. Dataset

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. There are 8 independent variables in this dataset, they are

pregnant: Number of times pregnant

glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test

diastolic: Diastolic blood pressure (mm Hg)

triceps: Triceps skinfold thickness (mm)

insulin: 2-Hour serum insulin (mu U/ml)

bmi: Body mass index (weight in kg/(height in metres squared))

diabetes: Diabetes pedigree function

age: Age (years)

And there is 1 dependent variable in this dataset, it is

test: test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

So the A matrix is a 768*8 matrix, and the b vector is 768*1 vector. This is an overdetermined project.

2. Objectives

1). To investigate factors which are related to diabetes.

2). Design a classifier to predict diabetes.

3). To evaluate the accuracy of the prediction.

4). Use logistics regression to fit the data, and make prediction.

5). To evaluate the accuracy of the prediction by logistic regression.

6). Compare the classifier prediction with the logistics regression prediction, choose the best prediction, and figure out the reason.

3. Methods

3.1. Feature selection

There are 8 independent variables (i.e., features, factors) in the diabetes dataset. We will use Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) to investigate which factors seem to be most import. Based on the result, we may sift out several most important factors to train the classifiers (i.e., Least Squares, Logistic Regression, K Nearest Neighbor). The performance of so-trained classifiers will be compared against the performance of the classifiers trained with all the factors involved. We expect to find a subset of the factor that gives acceptable classification accuracy.

3.2. Least squares classifier

We will start with Least Squares (LS) classification of the diabetes dataset by solving $\min_x||Ax - b||$. Here A and b are constructed using training samples. This will provide a offset and a set of weights on the factors. Then these weights can be used to classify held-aside test samples. By nature, LS classifier is a linear model which is equivalent to fitting a first-order polynomial to the training samples. The performance of LS classifier will be compared against other non-linear classifiers (e.g., Logistic Regression, KNN).

3.3. *Logistic regression*

Logistic regression is commonly used in discrete data analysis. It is used for predicting binary dependent variables. Formally, the logistic regression model is

$$log\frac{p(x)}{1-p(x)} = \beta 0 + \beta' x$$

Where $\beta 0\ and\ \beta$ are two parameters we can get by training the data we have known. We can choose w which maximize the likelihood function L :

$$[\beta 0, \beta] = arg\ max\ L(\beta 0, \beta)\ = arg\ max\ \Pi p(xi)^{yi}(1 - p(xi))^{1-yi} =$$

Xi is a vector of feature we observed, yi is an observed class (1 or 0).

Logistic regression allows us apply Logistic regression to our dataset to do a binary classification (diabetes or not), we can calculate

$$f(x) = log\frac{p(x)}{1-p(x)} = \beta 0 + \beta' x$$

if f(x) is non-negative we guess 1, otherwise, we guess 0.

Logistic regression is a kind of generalized linear model whose output has a Bernoulli distribution. For this dataset, we will use logistic regression model, but we may use other generalized linear model in the future to analyze other related dataset.

3.4. *K-Nearest neighbor*

Except for linear method, some of non-linear methods can be used for our data. For example, k Nearest Neighbor (kNN) algorithm. In this case, when an unclassified input is introduced, its label (1 or 0) will be determined by the k closest training examples (which label has more training examples).

KNN and some other non-linear classification method will be used to compare with linear classification methods.

3.5. Evaluation

There are 768 set of data in total. Cross validation will be used to evaluate the performance of each methods. 668 set of data will be randomly chosen for training, the rest of 100 set of data will be used to calculate the the accuracy. Repeat cross validation for 10 times for each method. Then calculate the mean of accuracy for each method to evaluate which method is most accurate.