Pima Data Set

This study is about Pima Indians, a community of American Indians in Arizona, which have an unusual high incidence and prevalence of diabetes mellitus. The data set has 9 variables, (see Table below), and the typical machine learning task is to predict the class variable of whether the subject has been diagnosed as diabetic or not.

1. Number of times pregnant [npreg]

2. Plasma glucose concentration [glu]

3. Diastolic blood pressure (mm Hg) [bp]

4. Triceps skin fold thickness (mm) [skin]

5. [2-Hour serum insulin (mu U/ml)]

6. Body mass index (weight in kg/(height in m) 2 ) [bmi]

7. Diabetes pedigree function[ped]

8. Age (years)[age]

9. Diabetes (No or Yes) [Type]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table : Variables in the Pima data set. From description at UCI Machine Learning Repository.

Following the book entitled :

Data Mining with Python (Working draft)

Finn Årup Nielsen

November 29, 2017

Here Date Set is build taken the results of insulin test each 2-Hours serum insulin (mu U/m), and by consequence this variable numbered 5 doesn’t appear at the following analysis

Data Set

pima\_training: 200\*8

pima\_testing :332\*8

**Logistic Regression**