Machine Learning and Pattern Recognition LAB

LAB 3: Dimension Reduction Problem + Bayesian Classifier

Predict the Onset of Diabetes

The test problem we will use in this tutorial is the Pima Indians Diabetes problem.
Download it from : https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f
This problem is comprised of 768 observations of medical details for Pima indians patents. The records describe instantaneous measurements taken from the patient such as their age, the number of times pregnant and blood workup. All patients are women aged 21 or older. All attributes are numeric, and their units vary from attribute to attribute.

# 1. Number of times pregnant
# 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test # 3. Diastolic blood pressure (mm Hg)
# 4. Triceps skin fold thickness (mm)
# 5. 2-Hour serum insulin (mu U/ml)
# 6. Body mass index (weight in kg/(height in m)^2)
# 7. Diabetes pedigree function
# 8. Age (years)
# 9. Class variable (0 or 1)

Write a code for

1. a) Principal Component Analysis (PCA) to the Pima Indians Diabetes.
2. b) Then apply Naive Bayes Algorithm

Observation /Comments:

1. There are 9 features:
   a. Number of times pregnant
   b. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
   c. Diastolic blood pressure (mm Hg)
   d. Triceps skin fold thickness (mm)
   e. 2-Hour serum insulin (mu U/ml)
   f. Body mass index (weight in kg/(height in m)^2)
   g. Diabetes pedigree function
   h. Age (years)
   i. Class variable (0 or 1)
2. There are two classes: 0 and 1
3. There are 768 records in this dataset
4. PCA is applied to reduce the dimensions. Then naïve bayes is applied. Accuracy of naïve bayes on dataset before and after PCA is calculated and compared.

```
accuracy of the model after PCA is 0.8311688311688312
accuracy of the model before PCA is 0.8051948051948052
```