Machine Learning: An Introduction
Assignment Problem #1

Prof. Philip Ogunbona
OAU Machine Learning Study Group
Obafemi Awolowo University, Ile-Ife Nigeria

Due date: To be discussed in class

Motivation

The goal of this assignment is to design and compare a number of classifiers using the data set provided. To facilitate this learning process, the assignment has been set up as a competition. We are seeking the best classification (or prediction) accuracy rate obtainable by any group in the class and based on any classifier. The choice of classifiers is not restricted. Each group will have to study the data carefully by reading about the features (variables), particularly the range of plausible values, meaning, method of measurement, etc. It is expected that a good deal of effort will need to be expended on data preparation (scaling, imputation, dealing with imbalanced data, etc.). The Machine Learning/Python books provided on Google Drive will be of great help in this regard. These books could also be used as de facto reference manual for Python modules (ScikitLearn, matplotlib, numpy, scipy, etc.) for Machine Learning. You should refer to the books on Machine Learning (also provided on Google Drive) for the theory underlying the various classifiers that you may choose to use in your experiment. You do not need to limit yourself to the books on Google Drive. Consult any resource you can access.

About the data

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of gathering this data was to design a classifier (or predictor) that will use the diagnostic measurements (features) and classify a subject (person) as having/not having diabetes. All subjects in the dataset are females at least 21 years old of Pima Indian heritage.

Features/variables The dataset is organised such each row contains the features for a subject. The columns contain the following features:

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

1. (125 Marks) **Task**

The following should be taken as the specifications of this assignment.

1. Form a group of no more than 5 and no less than 4 to work on this assignment. Give your group a nice name. Send me the names of members of your group (philip.ogunbona@gmail.com).
2. Select at least four different classification methods and design a classifier using the dataset provided.
3. Use the sklearn module from Python 3.XX (scikit-learn.org/stable/index.html) and any other module that will make your programming life easy.
4. Report your best classification accuracy regularly on OAU-ML-SG forum.
5. Submit a six-page report on your results for grading. See specifications of the report below. Based on the marks awarded to each section of the report, it is hoped that teams will allocate their efforts.

It is advisable that each team divides the work effort in such a way that everyone in the team has opportunity to deepen their theoretical and programming skills in Machine Learning.

Report

Your report should be strictly according to the following format (i.e. headings):

Title (5 marks) - Give your report a nice title and write the names of the members of your team as well as their student numbers.

Introduction (10 marks) - Describe the data in your own words and highlight various statistics (mean, variance, etc.) along with any significant observation that could be gleaned from the data.

In this section, you should discuss the method you adopted to select your classifiers. Justify your choices.

Data preparation (20 marks)- Describe the various methods and implications of the data preparations you undertook. Note that this is very important as it would have significant impact on the accuracy obtained from your classifier. You should discuss how you split the data for training, validation and testing.

Classifiers (40 marks) - Describe the various classifiers you have tested in your experimentation. This is very important because it shows how well you understand the properties of the classifier you have chosen. It is expected that you will write equations that describe the classifier model.

Evaluation (20 marks) - Describe and justify the methods of performance evaluation you have adopted. State the comparative evaluation estimates and justify the differences. There is a pool of 20 **extra bonus** marks to be shared by the winning teams. This implies that if there are several teams that obtain similar winning accuracy, they will share 20 bonus marks. If there is only one winning team, a bonus mark of 7 is awarded to the team.

Conclusions (30 marks) - You are required to reflect and write about the differences amongst the various classifier models relative to their parameters, amount of data required for training, nature/format of data required and the accuracy obtained. In addition, you are required to reflect and describe any significant trend/observation you discovered with regards to what features may be dominant in determining whether a subject will have diabetes. For example, is there a subgroup of subject that are more likely to have diabetes?

What needs to be submitted?

PLEASE READ VERY CAREFULLY

You are required to submit your 6-page report according to the format specified above. The report should be typed (or typeset using LaTeX) with 11-point font, one-and-half spacing and 1.5 cm all round margin. Submitted report MUST be a PDF file. Any WORD document should have been converted to PDF before submission. Non-PDF reports will not be marked.

You must submit the code for all the classifiers used in your experimentation. You must prepare the code that provided the best accuracy in such a way that it can be tested easily with new and previously unseen data.

In addition you must archive or "zip" or "rar" your source code and submit along with your report. Place your report and your archived source code in a folder with your group name and and "zip" or "rar" the folder before submission.

The most popular programming language used in industry for machine learning is currently Python. You must use Python for this assignment. As previously indicated, you should use Python 3.xx and jupyter notebook for your machine learning studies. This will facilitate easy sharing of codes and will allow the marker to be able to run your code easily. Ensure that your code includes documentation and reasonable naming convention for variables and functions. This is worth a bonus 5 marks!

Only one submission is expected from each group.

The due date for this assignment will be negotiated in the class. It is envisaged that about 4 weeks will be sufficient for this assignment.