

CSCI933: Machine Learning - Algorithms and Applications

Assignment #2

Due date: 23:59 (Beijing Time), 15th May, 2024

Motivation

The goal of this assignment is to design and compare a number of classifiers using the data set provided. Each student should 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, etc.). The Machine Learning/Python book will be help in this regard. The book could also be used as the reference manual for Python modules (ScikitLearn, matplotlib, numpy, scipy, etc.) for Machine Learning. You should refer to the books on Machine Learning for the theory underlying the various classifiers that you may choose to use in your experiment.

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 organized 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. Blood Pressure: Diastolic blood pressure (mm Hg)
4. Skin Thickness: Triceps skin fold thickness (mm)
5. Insulin: 2-Hour serum insulin ($\mu\text{U/ml}$)
6. BMI: Body mass index ($\text{weight in kg}/(\text{height in m})^2$)
7. Diabetes Pedigree Function: Diabetes pedigree function
8. Age: Age (years)
9. Outcome: Class variable (0 or 1)

Task

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

1. This assignment is an **individual** task. Each student must finish the report and coding work individually.

2. Select at least **three** different classification methods from the candidates and design the classifiers using the dataset provided. The candidates include: Bayes Decision Rule, Logistic Regression, Perceptron Criterion, Support Vector Machines, k-Nearest-Neighbors.
3. Submit a **five-pages** report on your results for grading. See specifications of the report below. Based on the marks awarded to each section of the report. **Your report will be checked by Turnitin system on Moodle for plagiarism detection.**
4. Submit the **codes** for all the classifiers used in your experimentation.
5. Each student must submit your own submission via Moodle.

Report

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

Title (5 marks) - Give your report a nice title and write the name as well as your UOW 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 (35 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.

Conclusions (10 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 5-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. Specify your **UOW** student number on the report.

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 “zip” your source code and submit along with your report. Place your report and your archived source code in a folder with your name and “zip” the folder before submission. This is **important** because of the way Moodle submission works.

The most popular programming language used in industry for machine learning is currently Python. You should use Python for this assignment. As previously indicated, you should use Python 3.xx for your machine learning studies. This will facilitate easy sharing of codes and will allow the markers 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!](#)