Machine Learning and Neural Computing: Data Classification Coursework

Date to be handed in: by 12pm on 28/03/2025

Introduction

In this coursework, you are required to produce an experimental report. Your tasks include analysing the data structure using principal component analysis (PCA) and building support vector machine (SVM) classification models to diagnose diabetes. The type of SVM you need to use is the C-SVC (Cost-Support Vector Classifier), and the kernel function you should use is the Gaussian radial basis function (RBF).

The data provided are extracted from the Diabetes Health Indicators dataset[1]. The data were divided into two groups: participants who had not engaged in physical exercise during the past 30 days were classified as the *NoActivity* group; all others were classified as the *PhysActivity* group. The data were further divided within each group into the training and test sets. Therefore, there are four datasets, namely, *diabetes_NoActivity_training.csv, diabetes_NoActivity_test.csv, diabetes_PhysActivity_training.csv* and *diabetes_PhysActivity_training.csv*. You can access these four data files from the module site on Canvas.

These datasets are balanced, each containing two classes: 0 (no diabetes) and 1 (prediabetes or diabetes). Each training dataset consists of 700 items, with 348 labeled as no diabetes and 352 as prediabetes or diabetes. Each test dataset consists of 300 items, with 152 labeled as no diabetes and 148 as prediabetes or diabetes. You can assume that the data is of satisfactory quality and requires no preprocessing/data cleansing other than normalisation.

Each row in the files, except for the header row, represents one participant. Each data file has 8 columns, where the first column is *Diabetes_binary*, that is, the class label, and all other 7 columns are features. Details about each column are provided in the Appendix.

[1]The dataset information can be found at `https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators`. The complete dataset can be found at `https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset`

Software Required

For this coursework, you will need to write your Python code (in version 3 and above) in the Jupyter Notebook. You can use functions from the following packages: Numpy, Pandas, Matplotlib, Seaborn, and Sklearn. Your practical session notes should be useful - these are all available on *Canvas*.

Tasks

1. Task 1 - Data Exploration (35 marks)
   In this task, you need to use Principal Component Analysis (PCA) to understand the characteristics of the datasets.

   (a) Use Pandas to load all four data files. For each dataframe, save 7 features and the label column into two different variables. You should explain which Python functions are used to do the task and the values you have chosen for them, even if you use the default values. (6 marks)

   (b) Produce a scatter plot comparing two features for each dataset. You may choose which two features to use, but your selection must remain consistent across all four datasets. Display all four scatter plots in a single figure arranged in two rows and two columns. You need to set the label for the x-axis and y-axis separately and use different colours to distinguish the classes. In your report, present the figure, describe the plots in your own words, and write your findings. (9 marks)
   Hint: examples on how to use pyplot.subplot in matplotlib can be found from the following link:
   `https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplot.html`

   (c) Normalise the training set (which is denoted as the training set (I)) and the test set separately for each group (that is, the *NoActivity* group and *PhysActivity* group) using **StandardScaler()** from **sklearn**. Explain in your report how you have normalised the datasets and report the mean and standard deviation for the first feature in the normalised test sets. (6 marks)

   (d) Perform a PCA analysis separately for each scaled training set. In your report, include a table showing the percentage of variance explained by each principal component for each PCA analysis. In addition, display two PCA visualization figures using the first principal component (PC1) and the second principal component (PC2). In each figure, label the data using different colours according to its class. Ensure taht the x-axis and y-axis are labeled appropriately. Describe each figure in your own words and write your findings. (14 marks)

2. Task 2 - SVM Classification on the *NoActivity* Group (18 marks)

   This task involves building a support vector classifier using **SVC** from **sklearn** library. Follow the steps below, and in your report, describe in your own words what you did and the results you obtained.

   (a) Data Preparation (5 marks)

i. Divide the training dataset into a smaller training set (II) and a validation set using the **train_test_split** function from **sklearn** and report the number of points in each set. Usually, we use 20%-30% of the total data points in the whole training set as the validation data. It is your choice on how to set the exact ratio.

ii. Normalise both the training set (II) and the validation set.

In your report, write which ratio you use for splitting, explain how you have normalised the two sets, and report whether, in general, the mean value and the standard deviation of each feature in the normalised training set are equal to the mean and standard deviation of the corresponding feature in the normalised validation set.

(b) Demonstrate how the models vary in performance with three different combinations of parameters $[C, \gamma]$: $[1, 1]$, $[5, 0.5]$, and $[0.5, 0.05]$ when using kernel **rbf**. (6 marks)

(c) Explain in your report how you have determined which combination of parameter values to use from the results you obtained in Task 2 (b). (2 marks)

(d) You should now be able to further test your model with the selected parameter values by classifying the test data. Present the results in your report, including the accuracy rate and the confusion matrix, in your own words. Avoid copying outputs or definitions, and provide a clear interpretation of each value, explaining what they reveal about your model's performance. (5 marks)

Hint: You will need to train an SVM model with the suitable parameter values discovered in Task 2 (c) on the whole normalised training set (I).

3. Task 3 - SVM Classification on the *PhysActivity* Group (18 marks)
Carry out Steps (a)-(d) shown in Task 2 on the training set and test set belonging to the *PhysActivity* group.

4. Task 4 - Cross-model Check (29 marks)

(a) Evaluate the *NoActivty* group model: Test the SVM model trained on the *NoActivity* group dataset using the normalised *PhysActivity* group test set.

(b) Evaluate the *PhysActivty* group model: Test the SVM model trained on the *PhysActivity* group dataset using the normalised *NoActivity* group test set.

(c) In your report, for each of the two sub-tasks above:

i. Clearly explain how each test set is normalised, specifying which training set's mean and standard deviation were used for normalisation. (4 marks)

ii. Clarify which trained model, obtained in which step of the previous tasks, is used and explain why it was chosen. (4 marks)

iii. Present the results, including information on the accuracy rate and the confusion matrix, in your own words. Provide a clear interpretation of each value, explaining what they reveal about your model's performance. (10 marks)

(d) Conclusions (11 marks)
Compare the results of both cross-model tests. Address the following questions in your report:

i. Which model performs better on the other group's dataset? Or are their performances similar? Why do you think this is the case? What evidence supports your view? (6 marks)

ii. Are there specific features or patterns in one group that might make it challenging for a model trained on the other group to generalise? (5 marks)

What to Submit

- The deliverable for this coursework includes
  - an experimental report with no more than ten pages. Please use a single-column format. The font size should be set to 11 or 12 points, and the line spacing should be set to 1.5 lines or single) in PDF format.
  - a Jupyter Notebook including all code you have written for this coursework.

  You must submit both files. The experimental report in PDF format will receive the plagiarism review via *Turnitin*; the Jupyter Notebook will be used to check whether the code works. No marks will be awarded if only one of the files is submitted.
- The structure of the experimental report
  - For each task, you may write a section to describe what you did, your results, and your findings in your own words. Showing the splitted training, validation, and test sets is unnecessary. Screenshots are typically not required. Do not screenshot/copy Jupyter Notebook's output directly.
  - This is an experimental report, and it is not necessary to write a literature review in the report, so references are not expected.

Please name both submissions using your student ID. For example, 17000000.pdf and 17000000.ipynb.

The total score of this coursework is 100, which is worth 30% of the overall assessment for this module. Note that you must do this coursework individually. You need to submit your coursework via Canvas to the assignment portal: Data Classification.

Please note that the 'Turnitin Submission for Data Classification' portal is for you to obtain a text-matching similarity report to improve your academic writing as necessary. You can submit your work to Turnitin once. Please do not submit the work (that is your final submission) that you would like to be marked there.

Guidelines

In this assessment you are permitted to use genAI tools (or a proofreader or proofreading service) to proofread your work but not permitted to use AI tools in the creation of content for your work. To do so would be considered to be academic misconduct. To what extent can you use genAI tools or a proofreader or a proofreading service to help you with your assessment? Neither a proof-reader nor a proof-reading tool (whether genAI or not) can ever be used to make changes to your work directly; the proof-reader or proof-reading tool must only identify and draw attention to possible changes which you can then choose to accept or reject; this will ensure that you remain the author

of your work. For clarity, where a proofreader, proofreading service or genAI tool is used, they/it may only:

- identify spelling and typographical errors;
- identify poor grammar;
- highlight formatting errors or inconsistencies;
- identify errors in labelling of diagrams, charts or figures;
- identify areas for possible improvement;
- highlight a sentence or paragraph where the meaning is not clear; or draw attention to repeated phrases or omitted words.

If you use a proof-reading service, which includes an AI Tool (e.g. Grammarly) you must declare this, otherwise this would also be an academic misconduct offence under the UPR 14 App3.

Appendix A Attribute Information

The following lists the attribute names and descriptions of the survey questions for participants, which is obtained from `https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators`.

1. BMI: Body Mass Index.
2. GenHlth: Would you say that in general your health is: scale 1-5
   1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.
3. MentHlth: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days.
4. PhysHlth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days.
5. Age:13-level age category (_AGEG5YR see codebook)
   1 = 18-24, 9 = 60-64, 13 = 80 or older.
6. Education: Education level (EDUCA see codebook) scale 1-6
   1 = Never attended school or only kindergarten, 2 = Grades 1 through 8 (Elementary), 3 = Grades 9 through 11 (Some high school), 4 = Grade 12 or GED (High school graduate), 5 = College 1 year to 3 years (Some college or technical school), 6 = College 4 years or more (College graduate).
7. Income: Income scale (INCOME2 see codebook) scale 1-8
   1 = less than $10,000, 5 = less than $35,000, 8 = $75,000 or more.