Assignment 2.: Multi-class classification of English Words using Support Vector Machine (SVM)

Dr. Suyong Eum

1^{st} Semester, 2023

1 Description

The aim of this assignment is to verify and to consolidate your understanding of Support Vector Machine (SVM) and Principal Component Analysis (PCA), which are fundamental tools in traditional[1] machine learning methods. SVM and PCA have been widely used in many practical problems.

In this assignment, you are expected to classify English words in terms of their difficult levels. You are given a data set which includes 11,999 English words which has 12 classes: 1(easy) to 12(difficult). In the file holding the data set, there are two columns: difficulty level and its corresponding English word. Your task is to create a model which defines the difficulty level of a given English word using Support Vector Machine (SVM).

You need to first define features which capture the difficulty level of individual English words such as the number of characters and word frequencies in a corpus, etc., under the assumption that people perceive a lengthy English word more difficult. Then, using the Principal Component Analysis (PCA), you analyze the importance of individual features you defined. Also, you are required to visualize the data set in two or three dimension figure through the dimensionality reduction provided by PCA.

2 Required Tasks

1. Please, try different kernel functions and parameter values to improve the accuracy of the classification result.
2. PCA is a great tool to verify your selection of features and also to visualize the data set. Please, visualize the data in 2D or 3D space.
3. You need to submit
   (a) A report which explains the model you developed including its accuracy.
   (b) A code: python or Jupyter notebook file.

3 Administrative

- Due: 24:00, July 12, 2023
- The data file can be downloaded from (www.suyongeum.com/B5G6G/)
- Submission to (suyong@ist.osaka-u.ac.jp)
  - Please, zip the code and report. Then, name it with your student number and assignment number, e.g., 32A18041_2.zip
- Late submission will be penalized at the rate of 10% reduction per day

[1] Comparing to Deep Neural Networks (DNN)