Assignment 1.: Multi-class classification of English Words using Support Vector Machine (SVM)

Practical Machine Learning: Dr. Suyong Eum

1^{st} Semester, 2018

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. We provide the training data set only. Thus, you need to evaluate the accuracy of your proposed model and parameter set using Cross Validation (CV).

2 Required Tasks

The assignment can be completed individually or as a group of 3 or less. There will be no advantage or disadvantage in terms of the number of people for the completion of the assignment.

1. Please, try different kernel functions and parameter values to improve the accuracy of the classification result.
2. Also, be careful not to overfit the data set and so you may consider carrying out the Cross Validation (CV) of your model and provide its average accuracy.
3. 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.
4. You need to submit
   (a) A report which explains the model you developed including its accuracy.
   (b) A code: python or Jupyter notebook file.
5. Bonus: you will get 10% bonus if you compare SVM result to that of Deep Neural Network (DNN).

3 Administrative

- Due: 6pm, June 8, 2018
- The data file can be downloaded from (www.suyongeum.com/ML/assignments.php)
- Submission to (suyong@ist.osaka-u.ac.jp)
  - Zip the report and code as one file and name it with your student number
- Late submission will be penalized at the rate of 20% reduction per day

[1] Comparing to Deep Neural Networks (DNN)