NCTU Pattern Recognition, Homework 4

**Deadline: May 25, 23:59**

**Part. 1, Coding (50%)**:

In this coding assignment, you need to implement the cross-validation and grid search using only NumPy, then train the [SVM model from scikit-learn](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html) on the provided dataset and test the performance with testing data. Find the sample code and data on the GitHub page

<https://github.com/NCTU-VRDL/CS_AT0828/tree/main/HW4>

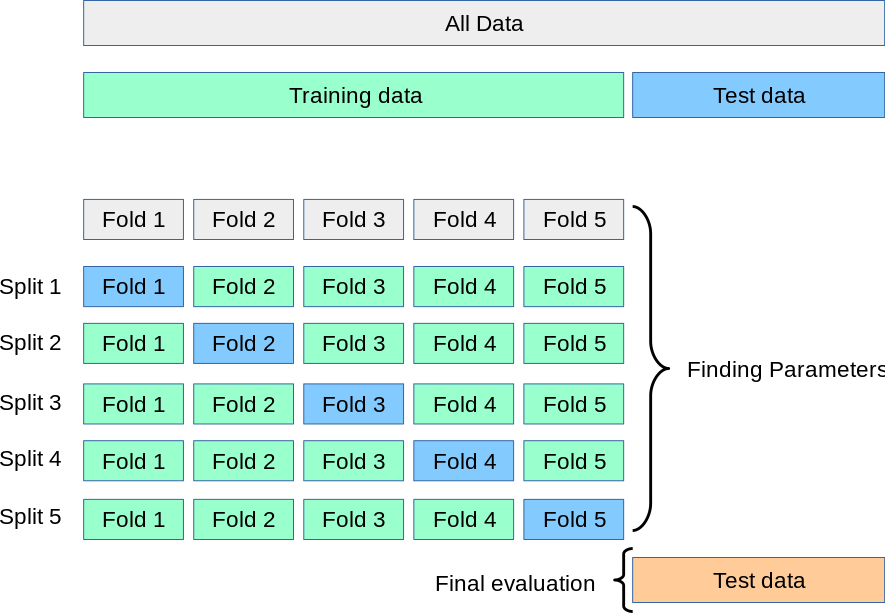
**Please note that only NumPy can be used to implement cross-validation and grid search. You will get no points by simply calling** [**sklearn.model\_selection.GridSearchCV**](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)**.**

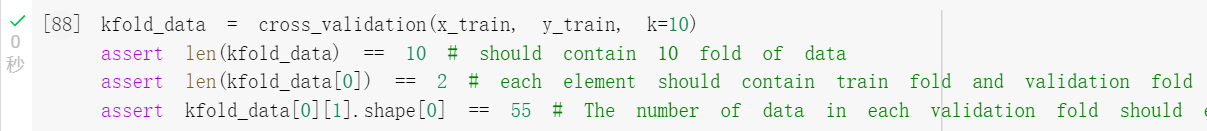
1. (10%) K-fold data partition: Implement the K-fold cross-validation function. Your function should take K as an argument and return a list of lists (*len(list) should equal to K*), which contains K elements. Each element is a list containing two parts, the first part contains the index of all training folds (index\_x\_train, index\_y\_train), e.g., Fold 2 to Fold 5 in split 1. The second part contains the index of the validation fold, e.g., Fold 1 in split 1 (index\_x\_val, index\_y\_val)

Note: You need to handle if the sample size is not divisible by K. Using the strategy from [sklearn](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html?highlight=k%20fold#sklearn.model_selection.KFold). The first n\_samples % n\_splits folds have size n\_samples // n\_splits + 1, other folds have size n\_samples // n\_splits, where n\_samples is the number of samples, n\_splits is K, % stands for modulus, // stands for integer division. See this [post](https://stackoverflow.com/questions/32305683/10-fold-cross-validation-with-sample-size-that-is-not-a-factor-of-10) for more details

Note: Each of the samples should be used **exactly once** as the validation data

Note: Please **shuffle** your data before partition





1. (20%) Grid Search & Cross-validation: using [sklearn.svm.SVC](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html) to train a classifier on the provided train set and conduct the grid search of “C” and “gamma,” “kernel’=’rbf’ to find the best hyperparameters by cross-validation. Print the best hyperparameters you found.

Note: I use k=3, since k=5 or k=10 have lower performance on testing data.

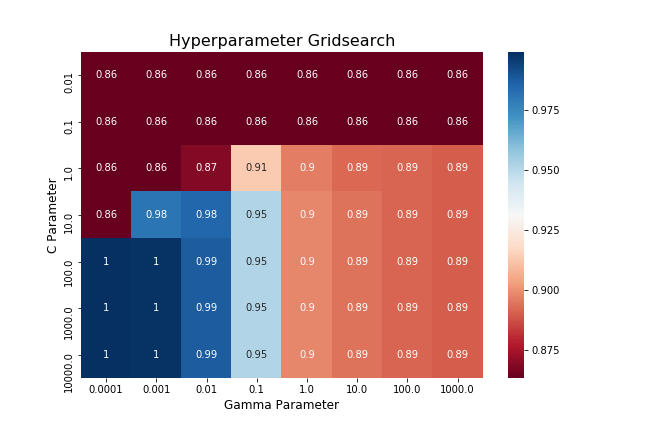
一張含有 文字 的圖片

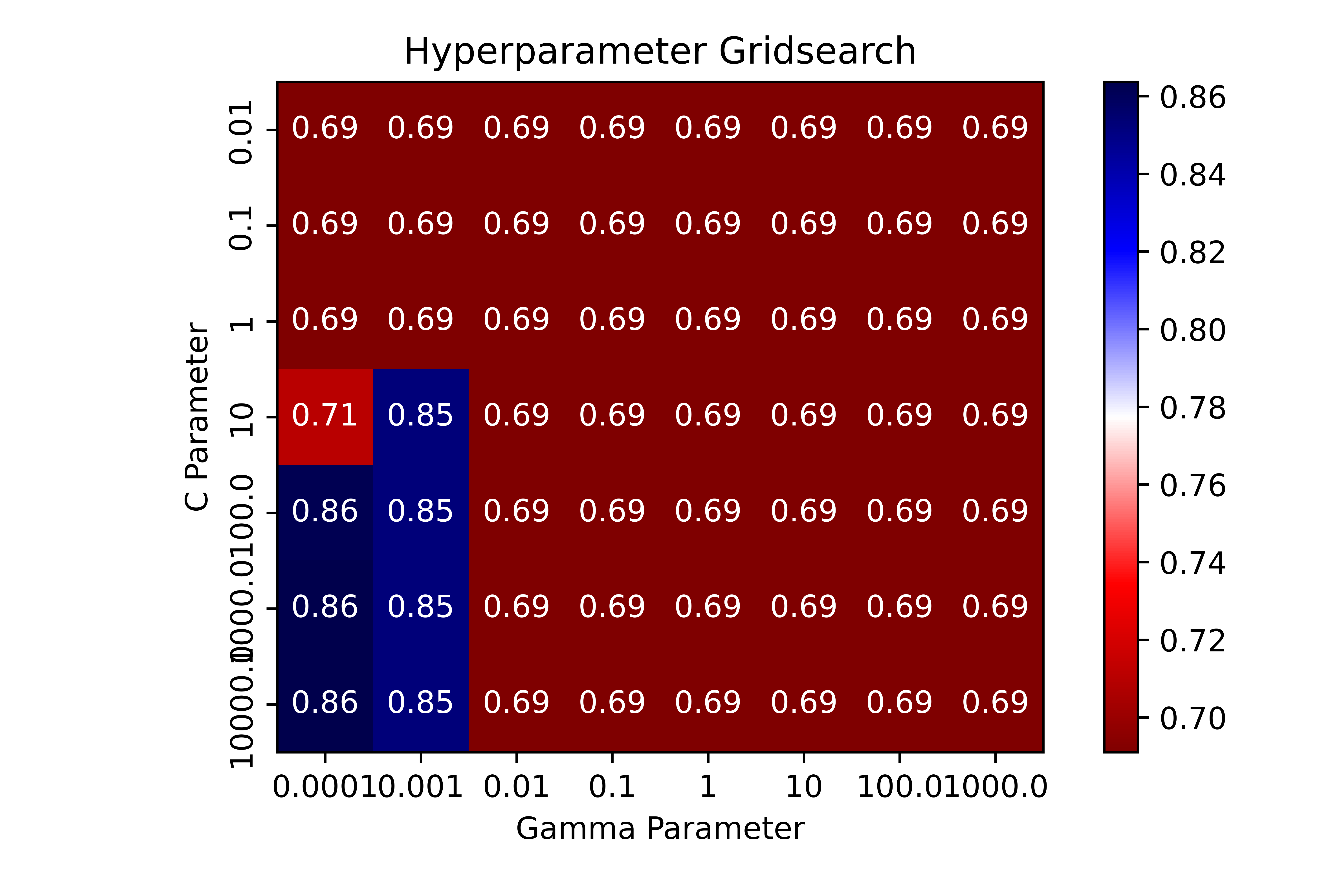
自動產生的描述

1. (10%) Plot the grid search results of your SVM. The x and y represent “gamma” and “C” hyperparameters, respectively. And the color represents the average score of validation folds.

*Note: This image is for reference, not the answer*

*Note:* [*matplotlib*](https://matplotlib.org/) *is allowed to use*





1. (10%) Train your SVM model by the best hyperparameters you found from question 2 on the whole training data and evaluate the performance on the test set.

|  |  |
| --- | --- |
| **Accuracy** | **Your scores** |
| **acc > 0.9** | **10points** |
| **0.85 <= acc <= 0.9** | **5 points** |
| **acc < 0.85** | **0 points** |

一張含有 文字 的圖片

自動產生的描述