NCTU Introduction to Machine Learning, Homework 4

Deadline: Nov. 29, 23:59

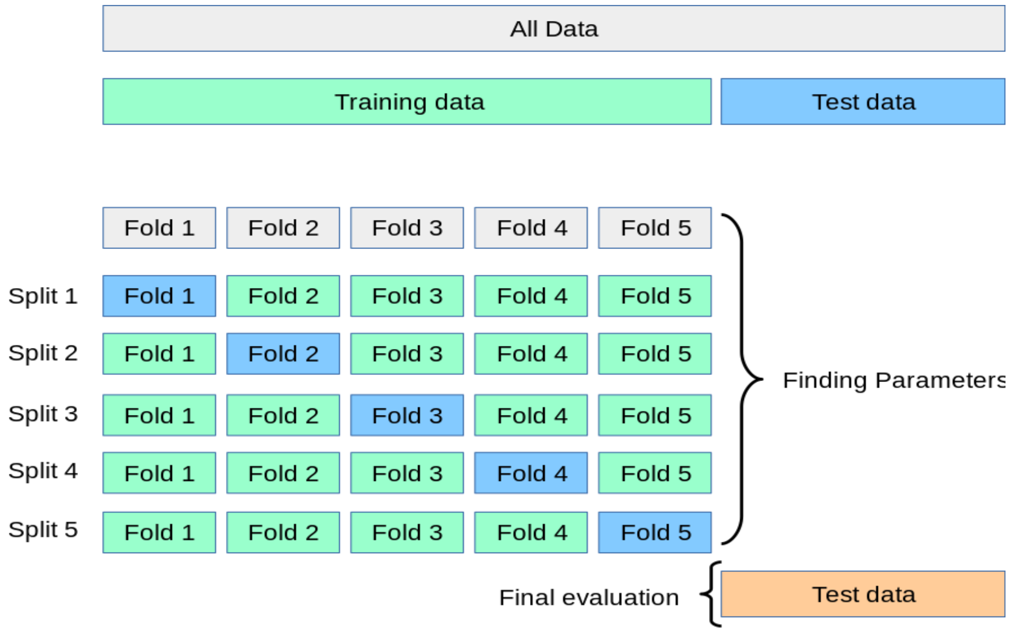
Part. 1, Coding (50%):

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 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_CS20024/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.

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)

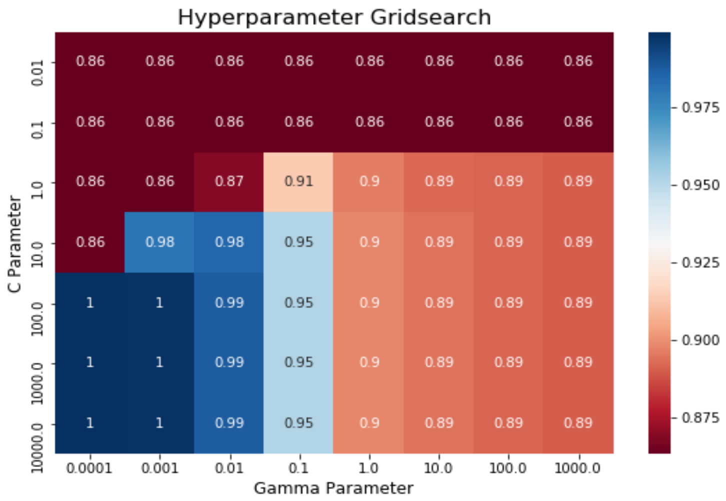
   Note: You need to handle if the sample size is not divisible by K. Using the strategy from sklearn. The first n_samples % n_splits folds have size n_samples // n_splits + 1, other folds have size n_samples // n_splits, where n_samples is the number of samples, n_splits is K, % stands for modulus, // stands for integer division. See this post for more details
   Note: Each of the samples should be used **exactly once** as the validation data
   Note: Please **shuffle** your data before partition



2. (20%) Grid Search & Cross-validation: using sklearn.svm.SVC 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: We suggest using K=5

3. (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 is allowed to use



4. (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

Part. 2, Questions (50%):

(10%) Show that the kernel matrix $K = \left[k\left(x_n, x_m\right)\right]_{nm}$ should be positive semidefinite is the necessary and sufficient condition for $k(x, x')$ to be a valid kernel.

(10%) Given a valid kernel $k_1(x, x')$, explain that $k\left(x, x'\right) = exp(k_1\left(x, x'\right))$ is also a valid kernel. Your answer may mention some terms like ____ series or ____ expansion.

(20%) Given a valid kernel $k_1(x, x')$, prove that the following proposed functions are or are not valid kernels. If one is not a valid kernel, give an example of $k(x, x')$ that the corresponding K is not positive semidefinite and show its eigenvalues.

a. $k\left(x, x'\right) = k_1\left(x, x'\right) + 1$

b. $k\left(x, x'\right) = k_1\left(x, x'\right) - 1$

c. $k\left(x, x'\right) = k_1\left(x, x'\right)^2 + exp\left(\|x\|^2\right) * exp(\|x'\|^2)$

d. $k\left(x, x'\right) = k_1\left(x, x'\right)^2 + exp\left(k_1\left(x, x'\right)\right) - 1$

(10%) Consider the optimization problem

$$minimize\ (x - 2)^2$$
$$subject\ to\ (x + 3)(x - 1) \leq 3$$

State the dual problem.