NYCU Pattern Recognition, Homework 4

Deadline: May 17, 23:59

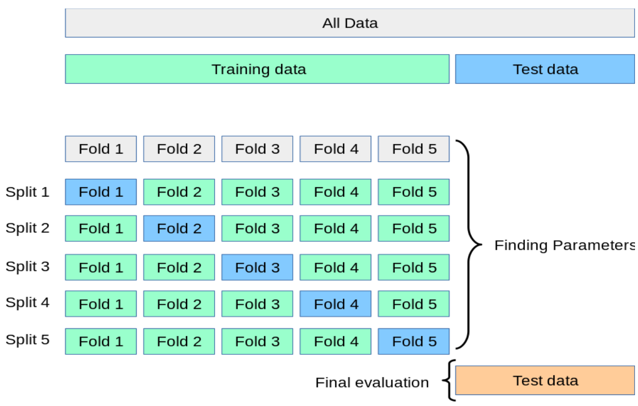
Part. 1, Coding (50%):

For this coding assignment, you are required to implement Cross-Validation and Grid Search using only NumPy. After that, you should train the SVM model from scikit-learn on the provided dataset and test the performance with the testing data. **You will get no points by simply calling sklearn.model_selection.GridSearchCV.**

(50%) K-Fold Cross-Validation & Grid Search

Requirements:

- Implement **K-Fold Cross-Validation** by creating a function that takes K as an argument and returns a list of K sublists.
    - Each sublist should contain two parts:
        - The first part contains the index of all training folds (index_x_train, index_y_train), for example, Fold 2 to Fold 5 in split 1.
        - The second part contains the index of the validation fold (index_x_val, index_y_val), for example, Fold 1 in split 1 .
    - You need to handle if the sample size is not divisible by K.
    - The first **n_samples % n_splits** folds should have a size of **n_samples // n_splits + 1**, and the other folds should have a size of **n_samples // n_splits**. Here, n_samples is the number of samples and n_splits is K.
    - Each of the samples should be used **exactly once** as the validation data.
    - Please **shuffle** your data before partition.

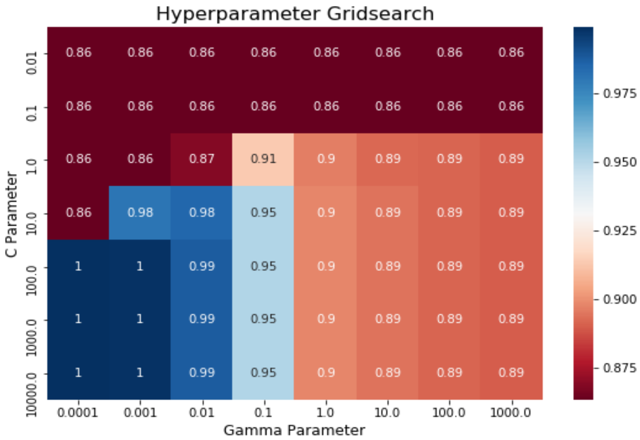


- Implement **Grid Search & Cross-Validation**:
    - Using sklearn.svm.SVC to train a classifier on the provided train set and perform **Grid Search** to find the best hyperparameters via cross-validation.

Criteria:

1. (10%) Implement K-fold data partitioning.
2. (10%) Set the kernel parameter to 'rbf' and do grid search on the hyperparameters **C** and **gamma** to find the best values through cross-validation. Print the best hyperparameters you found. Note that we suggest using K=5 for the cross-validation.

3. (10%) Plot the results of your SVM's grid search. Use "gamma" and "C" as the x and y axes, respectively, and represent the average validation score with color. Below image is just for reference.



4. (20%) Train your SVM model using the best hyperparameters found in Q2 on the entire training dataset, then evaluate its performance on the test set. Print your testing accuracy.

Points	Testing Accuracy
20 points	**acc > 0.9**
10 points	**0.85 <= acc <= 0.9**
0 points	**acc < 0.85**

Part. 2, Questions (50%):

1. (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.

2. (10%) Given a valid kernel k1(x,x'), explain that kx, x'= exp(k1x,x') is also a valid kernel. (Hint: Your answer may mention some terms like ____ series or ____ expansion.)

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

4. Consider the optimization problem

$$\mathit{minimize}\ (x - 2)^2$$
$$\mathit{subject\ to}\ (x + 4)(x - 1) \leq 3$$

State the dual problem. (Full points by completing the following equations)

$$L(x, \lambda) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$\nabla_x L(x, \lambda) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

when $\nabla_x L(x, \lambda) = 0$,

$$x = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$L(x, \lambda) = L(\lambda) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$