SIT384 Cyber security analytics

Pass Task 9.1P: Grid Search with Cross-Validation

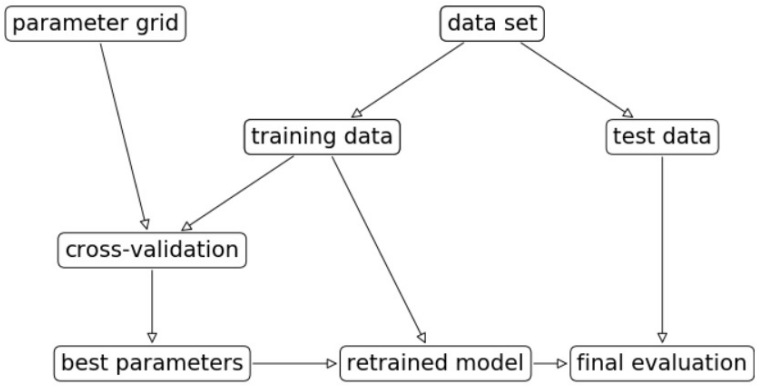
Task description:

C and Gamma are the parameters for a nonlinear support vector machine (SVM). The goal of SVM is to find a hyperplane that would leave the widest possible "cushion" between input points from two classes. There is a tradeoff between "narrow cushion, little / no mistakes" and "wide cushion, quite a few mistakes".

Small C makes the cost of misclassificaiton low ("soft margin"), thus allowing more of them for the sake of wider "cushion". Large C makes the cost of misclassification high ('hard margin"), thus forcing the algorithm to explain the input data stricter and potentially overfit.

The goal is to find the balance between "not too strict" and "not too loose". Gamma is the parameter of a Gaussian Kernel (to handle non-linear classification). Cross-validation and resampling, along with grid search, are good ways to finding the best C and gamma.

The following figure shows the process of finding the best parameters using grid search and cross-validation.



In this task, you are given a dataset, a parameter grid and cross-validation (CV) number, and try to find the best parameters C and gamma of SVM.

You are given:

- Dataset:
  from sklearn.datasets import load_digits
  digits = load_digits()
  X = digits.data
  y = digits.target
- Parameter grid:
  param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100],
  'gamma': [0.001, 0.01, 0.1, 1, 10, 100]}
- Other parameters of your setting

You are asked to:

- use Grid search with cross-validation to fit the data: SVC is the model, cv=5, return_train_score=True
- set random_state=0 when split train and test sets
- print grid_search.score of test dataset
- print grid_search.best_params_
- print grid_search.best_score_
- print grid_search.best_estimator_

Please refer to the textbook 5.2.3 Grid Search with Cross-Validation of chapter 5 “Model evaluation and improvement” and the textbook github site https://github.com/amueller/introduction_to_ml_with_python/blob/master/05-model-evaluation-and-improvement.ipynb (Grid search part).

Sample output as shown in the following figures are **for demonstration purposes only**. Yours might be different from the provided.

```
Parameter grid:
{'C': [0.001, 0.01, 0.1, 1, 10, 100], 'gamma': [0.001, 0.01, 0.1, 1, 10, 100]}
Test set score: 0.99
Best parameters: {'C': 10, 'gamma': 0.001}
Best cross-validation score: 0.99
Best estimator:
SVC(C=10, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma=0.001, kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. updated task9_1.py)
2. A screen shot of your program running

Check the following things before submitting:

1. Add proper comments to your code