A(E)3M33UI — Exercise E:
Training and testing error. Crossvalidation.

Petr Pošík

March 17, 2015

The goal of this exercise is to show how to correctly evaluate the quality of a predictive model using various validation methods.

1 Model validation

1.1 Having no testing data

As in the previous exercises, we shall work with the `auto-mpg.csv` dataset. Today we shall build a predictor of the car origin (USA or other country) based on all other car features.

Run the `exE.py` script. It shall load the data and print the shapes of `X` and `y` variables.

Task 1: In `exE.py`, fill in the code to train a SVM with RBF kernel (with default parameters) on the whole dataset, and measure its accuracy. Is it good? Should we use this model?

Hints:

- In previous exercises, we measured the model error by our own means, using the `model_evaluation.py` module. Now, we will use the built-in facility of scikit-learn, i.e. use the `score()` method of the SVM.

1.2 With training and testing data

To get at least some picture, how well the model will work on new data, let's split the data to training and testing data set, train it on the first and evaluate on both.

Task 2: In `exE.py`, fill in the code to split the data into training and testing data sets. Use the function `cross_validation.train_test_split`.

Hints:

- In the output of the scripts, check shapes of resulting Numpy arrays. Are they compatible?

Task 3: In `exE.py`, fill in the code to train the SVM on training data, and compute the accuracy on both, training and testing. How do they compare? Do you like what you see?

1.3 With *k*-fold crossvalidation

We would like to estimate the accuracy of the learning algorithm using crossvalidation.

Task 4: In `exE.py`, use the function `sklearn.cross_validation.cross_val_score` to get the CV estimate of the learning algorithm accuracy. You should get a list of the accuracies, one for each fold.

Hints:

- Look at the documentation for crossvalidation.

Task 5: Run the script several times. Do you see any fluctuations in the accuracy estimates based on the 3 above methods?

Task 6: How do you judge the SVM model? Is it properly set?

2 Model tuning

SVM models have several parameters that may be used to tune them.

Task 7: Find the meaning of `C` and `gamma` parameters of `sklearn.svm.SVC`

Hints:

- You shall have a general view from the SVM lecture.
- You may also find some info in the docs for kernel functions.

Task 8: Try to find a better setting for the `C` and `gamma` parameters of SVM by hand. What does "a better setting" actually mean?

2.1 Automatic tuning using grid search

Let's try to use grid search feature to search for optimal settings for SVM.

Task 9: Learn about the `GridSearchCV()` function in the documentation.

Task 10: Use `GridSearchCV()` to find near-optimal values of `C` and `gamma` for SVM with RBF kernel. *Use only the training part of data to search for the parameter values!*

Task 11: Print out the scores of the final classifier on training and testing data.

3 Conclusion

Dealing the datasets correctly with respect to the learning algorithms is a *crucial* thing in data analysis. Think about it thoroughly!

4 Have fun!

Complete the exercise as a homework, ask questions on the forum, and upload the solution via Upload system!