Workshop

Design and Analysis of Machine learning experiments

Junfeng Gao

UNIVERSITY OF LINCOLN

Aim

- The purpose of this workshop is to understand k-fold cross-validation (CV) and the confusion matrix. As introduced in the lecture, k-fold CV is the classical method for comparing machine learning algorithms as well as for selecting hyper-parameters.

- The confusion matrix is the table to visualize the performance of an algorithm. It is used to calculate the performance measures.

- You will use a small data set to practice k-fold CV and compute the performance measures.

Task

- 1) Download the data. The data has six columns as features and the last column is the class labels.

- 2) Apply random forests (RF) to this data using sklearn.ensemble.RandomForestClassifier. The numbers of trees (*m*) and max features (*n*) which must be chosen. The two parameters will influence the performance of RF. Therefore, you are asked to choose the optimal values for *m* and *n*. You could use a grid search method which was introduced in the lecture. To use cross validation (CV) you must choose a proper performance measure, we choose to use the accuracy.

- 3) You can use the tools in *sklearn, numpy* and *pandas* to accomplish these tasks. For example, you can use GridSearchCV function from sklearn.model_selection.

- 4) Report the accuracy you achieved and the optimal hyperparameters you selected.

Refer to the link below
https://scikit-learn.org/stable/modules/cross_validation.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html