sat sun mon tue wed thu fri Date :

Common Pitfalls and recommended practices

Speaking, scikit-learn data preprocessing and uncertainity are covered in this chapter. Inconsistent preprocessing causes model failure since train and test data should be scaled identically. Pipeline automation and dataloss minimization may overstate model training results. Choose ~~features~~ features from training date, not test. Managing uncertainity is hard. An integar for 'random_state' produces repeatable results, although cross-validation may be unreliable. RandomState instances test models in random situations to make them resilient. RandomForest and cross validation splitter yield mixed fold results based on "random_state". Avoid 'random_state=None' and follow CV unpredictability and RandomForestClassifier or other random estimator distribution guidelines.