Handson Machine Learning Notes

Part I

Supervised/UnSupervised/SemiSupervised/Reinforcement

Dimensionality Reduction/ Feature Extraction

Batch/Online Learning(train the system incrementally by feeding it data instances

sequentially)

Part II

Pipe line:A sequence of data processing components is called a data pipeline

strati‐ed sampling: the population is divided into homogeneous subgroups called strata, and the right number of instances is sampled from each stratum to guarantee that the test set is representative of the overall population

Replace NA value, median,delete,zero

Categorical-numerical(binary)

There are two common ways to get all attributes to have the same scale: min-max

scaling and standardization.

Cross-Validation

Part III Classification

The StratifiedKFold class performs stratified sampling (as explained in Chapter 2)

to produce folds that contain a representative ratio of each class. At each iteration the

code creates a clone of the classifier, trains that clone on the training folds, and makes

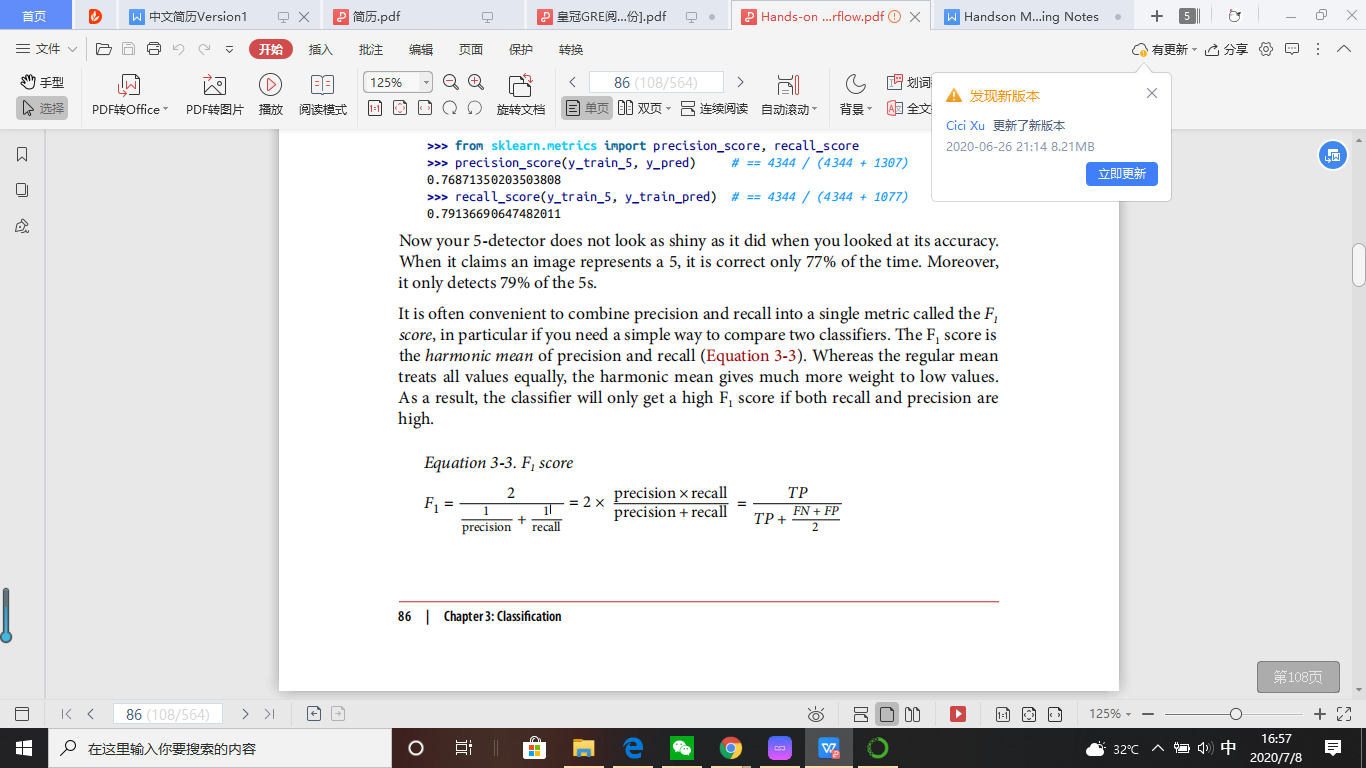
predictions on the test fold. Then it counts the number of correct predictions and

outputs the ratio of correct predictions.

1. fold crossvalidation means splitting the training set into K-folds, then making predictions and evaluating them on each fold using a model trained on the remaining folds

Accuracy is not a good matrix for skewed dataset

Confusion Matrix: the number of Class A that is recognized as Class B

precision = TP /TP + FP

TPR or sensitivity(recall) = TP /TP + FN

F1 score is the harmonic mean of precision and recall

Precision/recall tradeoff

receiver operating characteristic (ROC) curve plots the true positive rate (another name for recall) against the false positive rate. AUC: area under curve

one-versus-all (OvA) strategy: binary for one compare to all other class then combine

One-versus-one strategy: binary for Class A&B