Introduction to Machine Learning
Homework 2

October 15, 2021

Academic integrity

Our lesson cares much more on academic integrity. No matter who should do our utmost to handle the establishment of academic integrity standard including the host teacher and assistants of this lesson. We hope you will have the same faith with us.

(1) Discussion between students is allowing. The work named by yourself must be completed by your own hands. Any kind of Copying from existing documents will be seen as illegal.

(2) Any kind of Copying from other people's fruits of labour(Publication or Internet documents) will be accused of plagiarism. The score of plagiarists will be canceled. Please mark the authors if you cited any public documents of them;

(3) Highly resemble homework will be seen as Coping. No matter who you are, the one who copy or the one who is copied, both of your score will be canceled. Please protect your homework not to be copied by others actively.

Homework submission notes

(1) Please follow the submission methods on the website;

(2) If you are not follow the methods or your submission format are not correct. We will deduct some score of your homework;

(3) Unless some special cases, the submission over deadline will not be accepted and your score will be set as zero.

1 [20pts] Decision tree

Suppose you are given data consisting of a training set of 5 examples and a test set of 4 examples. Each sample in the training and test set has three binary features (A, B, C) and one binary label (y).

(1) Using the training set (Table 1), construct a decision tree for the binary classification. Use the Information Gain (IG) as the decision criterion to select which attribute to split on. Show your calculations for the IG for all possible attributes for every split. (If there are multiple optimal features when splitting, please select the feature with the smallest alphabetical order.)

(2) Now evaluate the decision tree you have created on the test set (Table 2).

A	B	C	y
1	0	1	1
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1

Table 1: Training set for decision tree

A	B	C	y
0	0	0	0
0	1	1	1
1	1	1	0
1	0	0	0

Table 2: Training set for decision tree

2 [30pts] Neural network

In this problem, you are asked to build a neural networks from scratch and examine performance of the network you just build on pendigits[1] data set. Here are some instructments listed below:

1. You are allowed to use out-of-the-box deep learning tools (e.g., PyTorch, TensorFlow, ...) to build your model.
2. You don't have to implement deep and complex neural networks to achieve the state-of-the-art performance. However, brief performance comparisons between different architectures, different hyper-parameters, and different optimization methods are needed.
3. You need to submit your code and describe how to use them. Briefly showing your analysis, experimental results, and conclusions in this homework is also necessary.

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/pendigits/

3 [50pts] Learn from imbalanced and noisy data

In real-world scenrios, data tends to exhibit a class-imbalanced distribution. Moreover, the labels annotated by workers are not always correct. Since it is too difficult to solve these problems in real-world data, we make a synthetic dataset based on pendigits. We corrupt the pendigits training data and generate its class-imbalanced and noisy version[2].

(1) Use at least two machine learning algorithms (e.g., Decision tree, Neural network, SVM, ...) to learn from the corrupted pendigits dataset, and validate on the original test dataset. Compare the performance and cost of different algorithms on this task.

(2) Analyze the impact of imbalance class distribution and label noise on the models. Note that, we generate the data under a specific imbalance ratio and noise level. You can try more settings by changing the value of 'imb_ratio' and 'noise_level' in **data_utils.py**.

(3) Provide at least one method to alleviate the class-imbalanced problem or the label noise problem. You can either give the theoretical viewpoint, or report the experimental results.

(4) Now we provide an anonymous dataset with class-imbalanced problem and label noise[3]. The formats of the training and test file are similar to pendigits, except that the labels of the test data are removed. You are asked to build a classifier with the training data, and predict on the test data. We will evaluate the accuracy of your prediction results with the test labels.

    Note that, the output of your classifier should be a txt file "output_yourId.txt" which contains the same number of lines as the test file, each line contains a single number. Please do not make confusion about the order of test example, otherwise you may get a very low accuracy.

You are allowed to use out-of-the-box machine learning tools (e.g., Scikit-learn, PyTorch, TensorFlow, ...) to build your model. For all experiments, you need to submit your code and describe how to use them.

[2] https://git.nju.edu.cn/shijx/pendigits-corrupted/
[3] https://box.nju.edu.cn/d/7b3790415acf49d08a34/