2AMM10 2021-2022 Assignment 4: Learning a distribution

1 Introduction

Learning the underlying distribution of a dataset can help with performing various tasks such as detecting out of distribution data points, learning relevant features of the data, and performing classification when only very few labels are present. This assignment serves as an example of that.

– 某些数据坏掉了
– 数据集中混入了一些其他数据集的数据
– 对于数据集中大部分的数据，我们不知道其所属的类

只有一少部分数据有label，其他数据label未知
数据集中的一大部分已经证明是非匿名的，也就是在这5个类中的

For this assignment, assume we have data coming from some process according to a distribution with five modes, i.e. the data can be said to belong to five classes. While collecting data from this process, something went wrong: some of the data got corrupted and moreover, some data that doesn't belong in the dataset accidentally got mixed up in it. Furthermore, for most of the data, we have no idea which of the five classes it belongs to. Because manually labeling data is expensive, only a small fragment of the collected dataset was labeled, and a slightly larger segment of the collected dataset has been verified to be without anomalies.

匿名，就是指属于数据集给出的class之外的class，我们不需要关心这个未知的class是什么class，只需要知道它不属于数据集的class

Your goal now is now to come up with a model that can be used for three things:

1. it should be able to detect out of distribution data; anomaly data
2. it should give a low (e.g. ten) dimensional description of the dataset in terms of the five modes of the distribution;
3. it should be able to classify the remaining data points into the five classes.

2 Data

In order for you to be able to have visual results, we have built the data sets for this assignment from FashionMNIST. This is however just for convenience, so you are not allowed to use a pre-trained model as you can not assume you have a similar enough dataset to pre-train a model on. However, we do want you to answer the following questions in the discussion section of your report:

1. In what situations would you be able to use transfer learning to aid you in the given tasks?
2. How would you use transfer learning to perform the given tasks? 到底可不可以用transfer learning? 这里只是理论论证吗
3. In what situations would you not be able to use transfer learning / when could transfer learning be detrimental to performance on (some of) the given tasks?

For this assignment, you are given access to four datasets:

– 一个数据集：26000个无label但是全部都属于数据集给出的class的数据
– 一个数据集：2000个有label且在数据集给出的class的数据
– 两个数据集：每个有1052个数据，全部有label，但是5%的数据是匿名的，也就是不属于5个class，将匿名的class标为第六个class

- a dataset containing 26000 data points which are known to be in-distribution (i.e. not anomalous) but which are without labels;
- a dataset containing 2000 labeled data points which are all in-distribution;
- two datasets containing 1052 data points each which contain roughly 5% out-of-distribution data and which are fully labeled (anomalies being labeled as a sixth class).

The data points consist of 32×32 pixel gray-scale images of datatype `uint8`, the labels of the first labeled dataset are one-hot encoded with five classes (datatype is `float32`), and the labels of the other two sets are one-hot encoded with six classes, (again of type `float32`).

The data can be found in `assignment_4_data.pickle`, which is pickled with protocol 4 (see `https://docs.python.org/3/library/pickle.html`). The pickled data is structured as

```
{
  'unlabeled_data': 'array',
  'labeled_data': {'data': 'array', 'labels': 'array'},
  'representative_set_1': {'data': 'array', 'labels': 'array'},
  'representative_set_2': {'data': 'array', 'labels': 'array'}
 }
```

with shapes

```
{
  'unlabeled_data': '(26000, 1, 32, 32)',
  'labeled_data': {'data': '(2000, 1, 32, 32)', 'labels': '(2000, 5)'},
  'representative_set_1': {'data': '(1052, 1, 32, 32)', 'labels': '(1052, 6)'},
  'representative_set_2': {'data': '(1052, 1, 32, 32)', 'labels': '(1052, 6)'}
}
```

Give a clear description of how you will use the various datasets to train and validate your model. Motivate your approach.

3 Deliverables

You should provide well documented Python code for both the creation, the training, and the evaluation of your model. For the creation and training of your model, use Pytorch. Moreover, you should provide a report in PDF format containing the following sections:

Problem formulation: Translate the described situation and given goals to a machine learning problem.

Model formulation: Develop a model that can be used to meet all three of the described goals. Describe how the model can be used to perform the relevant tasks. Motivate the design of your model in terms of the problem formulation, in terms of the availability of data, and in terms of data characteristics like symmetries and equivariances.

Implementation and training: Give a concise yet clear enough description of the implementation of your model and training procedure for someone unfamiliar with your code to be able to write their own implementation of your model and train it. Moreover, motivate your data augmentation, loss-functions, and training procedure in terms of your problem formulation, and data characteristics. In particular, describe how you use the available labels and how you use the unlabeled data.

Experiments and evaluation: Describe your evaluation strategy, perform experiments, and evaluate your model according to your strategy. Motivate your choice of metrics in terms of the given tasks.

Conclusion: Summarize your approach and results.

Discussion: Describe how your approach might be improved upon in the future given the same constraints as in this assignment. Hypothesize in what situations, and in what ways, transfer learning could aid in creating a model that can perform the given tasks.