Project #8
Adversarial Robustness in Machine Learning Models

Business Understanding Report

Members:
Nischay Uppal - 25%
Patrick Florendo - 20%
Supriya Dixit - 25%
Sai Coumar - 30%

Here we consider dataset(s) produced by the data preparation phase, used for modeling or for the major analysis work of the project.

1. SELECT DATA

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

1.1. Rationale for inclusion/exclusion

List the data to be used/excluded and the reasons for these decisions.

- Collect appropriate additional data (from different sources—in-house as well as externally) • Perform significance and correlation tests to decide if fields should be included
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data)
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experience of modeling (i.e., model assessment may show that other datasets are needed)
- Select different data subsets (e.g., different attributes, only data which meet certain conditions)
- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
- Document the rationale for inclusion/exclusion
- Check available techniques for sampling data

All data will be included. Running PCA on the images did not drastically improve performance in any metric because the images were already compressed (approximately 30 pixels x 30 pixels).

Additionally, because the only metadata available is the classification vectors, no data can be excluded without meaningfully affecting the trained models' output.

It should be noted that these datasets have been pruned from their initial sources. Specifically, MNIST discarded 50,000 images from NIST's original testing set, CIFAR10 went from 80 million images spanning 50,000 nouns to 60,000 images spanning 10 nouns, and SVHN has set aside 50,000 images as "extra data".

Due to the differing natures of each dataset (MNIST, CIFAR10, SVHN), we will be running tests on all of them to ensure our results are as robust and generalizable as possible. The MNIST dataset is valuable as a proof-of-concept, and CIFAR10 and SVHN allow us to experiment with more complex datasets.

The data has already been split into mutually exclusive subsets for training and testing purposes.

Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle

the weighting.

2. CLEAN DATA

Raise the data quality to the level required by the selected analysis techniques. This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.
Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results.

- Reconsider how to deal with any observed type of noise
- Correct, remove, or ignore noise
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer.
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).

We were not able to detect any deviation in the datasets that would be categorized as noise or outliers. Because these datasets have been pre-curated, our team did not have to do much image processing to clean the data.

Each dataset went through several stages of image processing to normalize the data. Originally, the training and testing data for the NIST database were collected from two different sources, so the two sources were combined before being split into new training and testing subsets for the MNIST database. Additionally, the images were compressed from 128x128 black-and-white images to 28x28 grayscale images.

The original dataset that CIFAR10 was drawn from, 80 Million Tiny Images, stored compressed, 32x32 images. The original, high-resolution versions of each image were never stored. During the creation of CIFAR10, new labels were generated instead of using the noisier tags from Tiny Images to ensure that there would be no overlap between classes.

The images in the SVHN dataset were pulled from Google Street View. Although the original images exist with bounding boxes for each digit, the format that our team is using has been standardized so that each digit is centered in a 32x32 image.

Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.

3. CONSTRUCT DATA

This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.

- Check available construction mechanisms with the list of tools suggested for the project • Decide whether it is best to perform the construction inside the tool or outside (i.e., which is more efficient, exact, repeatable)
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data construction (i.e., you may wish include/exclude other sets of data)

Our datasets come in different data types and require standardization into numpy arrays/pandas dataframes that could be reshaped into viewable images. Pandas has robust functionality to extract, transform, and aggregate attributes, and numpy arrays can be utilized to apply transformations to the data where it's necessary - such as pixel normalization.

MNIST:

MNIST data comes as a CSV of one-hot encoding vectors with one label column and 784 pixel data columns. This was the cleanest data format but viewing images was difficult to check samples. Functionality was implemented to strip the labels out and convert the flattened one-hot encoding vectors into 28x28 np arrays which can be viewed as images through visualization packages like PIL or matplotlib.

CIFAR10:

The CIFAR10 comes in batches of encoded data. Using pickle the data was decoded from bytes and appended into an accumulated numpy array. This data is also in the format of one hot encoding vectors and similarly needs to be unflattened into 32x32 arrays which can be viewed as images through visualization packages like PIL or matplotlib. One caveat was that these images had RGB channels and the flattened encoding vector had 1024 columns for the R values, 1024 values for the G values, and 1024 for the B values which needed to be processed and stacked to reassemble images.

SVHN

The SVHN data came as Matlab files and raw images. The raw images weren't labeled and were unprocessed so we used the Matlab file version which was loaded via scipy.io. This data came from large tensors with each image retaining its square shape which could be viewed. They needed to be converted back into standardized one-hot encoding vectors for certain summary statistics as well as reshaped for later use in modeling.

There is no empirical change in the selected attributes prior to modeling. Greyscaling and normalization is applied during model testing but not as an essential step of data construction.

All 3 datasets now have standardized flattened one hot encoding vectors and viewable images (and functionality to switch between these formats). All data can be freely converted between pandas dataframes and numpy arrays for whatever purpose is necessary, giving us a lot of freedom.

All datasets would later be reshaped into PyTorch Tensors for modeling. Rather than use flattened one-hot encoding vectors or the reshaped images with dimensions (height x width x # channels) PyTorch Image tensors are used (# channels x height x width). Therefore, all data needed to be flattened first and then reprocessed into that data format.

Data construction was done along with summary statistics in the following files:

1) mnist_summary_statistics.ipynb
2) cifar-10_summary_statistics.ipynb
3) svhn_summary_statistics.ipynb

3.1. Derived attributes

Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be: area = length * width.
Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources should be used in constructing a model. Derived attributes might be constructed because:

- Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it
- The modeling algorithm in use handles only certain types of data—for example we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model
- The outcome of the modeling phase suggests that certain facts are not being covered

Derived attributes

- Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate)
- [Optional] Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)
- How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).]
- Add new attributes to the accessed data

Derived attributes were not used given the nature of our simple image datasets. Some batch pixel normalization was applied in the modeling phase while testing models but were not applied universally or consistently while hyperparameter tuning.

No single-attribute changes were made. Image transformations (rotations, flipping, gaussian noise, cropping) were attempted as a part of initial data augmentation but led to overall decreased model performance. The well-curated nature of the dataset made excessive transformations to the data largely problematic.

CIFAR10 had string labels that were already mapped to numeric representations for modeling and a mapping back was added for human interpretability post-modeling/initial viewing of images.

3.2. Generated records

Generated records are completely new records, which add new knowledge or represent new data that is not otherwise represented (e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing).

- Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).

Not Applicable. Image Data cannot be used to generate records.

4. INTEGRATE DATA

These are methods for combining information from multiple tables or other information sources to create new records or values.
Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage, it may also be advisable to generate new records. It may also be recommended to generate aggregate values.
Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

- Check if integration facilities are able to integrate the input sources as required •

Integrate sources and store results

- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)

With respect to the image datasets themselves, they will not be combined in any way as the underlying distributions of each dataset is significantly different. Combining these datasets and training a model on it

will lead to subpar results as the model will learn characteristics from each dataset that are not necessarily transferable.

Remember that some knowledge may be contained in non-electronic format.

5. FORMAT DATA

Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.
Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Rearranging attributes

- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

As mentioned before, PyTorch tensors are in the shape of # channels x height x width per image rather than flattened one hot encoding vector or height x width x # channels. For MNIST and CIFAR10 the data was already in standardized one hot encoding vectors, but the SVHN data needed to be flattened from height x width x # channels into a one hot encoding vector. Then all three datasets were mapped to the # channels x height x width format.

Reordering records

- It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

No records reordering is required. Instead, it is recommended to shuffle the training dataset randomly.

Reformatted within-value

- These are purely syntactic changes made to satisfy the requirements of the specific modeling tool • Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)

All the attribute values are numeric values 0-255 (before normalization).

6. DATASET DESCRIPTION

Provide a general description of the final dataset (for instance, in terms of number of samples and number of features).

MNIST has 60,000 training samples and 10,000 test samples, with one-hot encoding vector format and image format available. There are 784 features for each one channel (pixels) and 1 label feature for a total of 785 features.

CIFAR10 has 50,000 training samples and 10,000 test samples, with one-hot encoding vector format and image format available. There are 1024 features for each color channel (pixels) and 1 label feature for a total of 3073 features.

SVHN has 73,257 training samples and 26,032 test samples, with one-hot encoding vector format and image format available. There are 1024 features for each color channel (pixels) and 1 label feature for a total of 3073 features.