L03: Classification

CSci 574 Machine Learning (Géron) Ch. 3

Derek Harter

Department of Computer Science
East Texas A&M University

Summer 2025

EAST TEXAS A&M

Introduction to the MNIST dataset

We will be using the MNIST dataset in our discussion of ML classification.

- MNIST is to ML as the "fruitfly" is to genetics
- MNIST Properties:
  - 70k black and white (single greyscale channel) images
  - Each a digit, so 10 distinct classes 0-9
  - Images are handwritten digit, 28x28 pixels

Fetching the digits using `sklearn.datasets` (returns a `dict`, there are other keys that may be useful):

```
from sklearn.datasets import fetch_openml
mnist = fetch_openml('mnist_784', parser='auto')
X, y = mnist['data'], mnist['target'].to_numpy()
```

```
X.shape
```

```
(70000, 784)
```

```
y.shape
```

```
(70000,)
```

EAST TEXAS A&M

Introduction to the MNIST dataset

Figure 1: Digits from the MNIST dataset. (Géron, 2023, pg.182)

For example, the first 100 digits

- Digits are already randomly shuffled, and dataset is fixed
  - So we can safely split off last 10k digits and reserve for testing

EAST TEXAS A&M

Binary Classification

- There are 2 fundamental types of supervised learning
  1. **Regression**: like predicting real valued house price from features
  2. **Classification**: for example, MNIST predict which digit among discrete labels 0-9
- The simpliest classification task is **Binary Classification**.
  - For example, let's use MNIST, but make a 5 / not-5 predictor

EAST TEXAS A&M

Measuring Performance of a Classifier

- Evaluating a classifier is significantly trickier than evaluating a regressor
- For example, for 5 / not-5 can use accurracy of correct predictions.
  - However, what is a good accuracy for this task?
  - 90% of the values are not-5, and 10% are 5, so always guessing not-5 gives 90% accuracy
- **Common Sense Baseline** You should always have a minimum common-sense baseline in mind when evaluating a ML system performance.
  - For example, when data is skewed, accuracy that doesn't do better than always guessing the most common class is not doing anything significant yet.

EAST TEXAS A&M

Bibliography

Géron, A. (2023). *Hands-on machine learning with scikit-learn, keras and tensorflow* (third). O'Reilly Media, Inc.

EAST TEXAS A&M