

You are a data scientist for the U.S. Postal Service. You are tasked with building an address detector so that mail can be automatically sorted. For this work, we will focus on hand-written digit recognition only.

The MNIST dataset of handwritten digits is extremely well studied¹. Each training example in this dataset is a single hand-written digit. The goal is to construct a classifier that can correctly identify the value of the digit from the image. The MNIST dataset comes with `scikit-learn` in the `datasets` subpackage. Take a look at documentation and online examples for information of how to load this dataset.

Building your classifier(s):

- You must implement and report on a simple case of an algorithm as described in class.
- You can use the `scikit-learn` library which is installed on all campus machines for a more robust analysis.

The Write-Up

You must present a model report to your company, providing the supervisors with an accurate depiction of your model, and how much they should trust your results. You can present results from more than one model, but you should include a recommendation to the company on which model they should use. *you need to be sufficiently precise with your writing and include enough detail that a competent reader could reproduce your results.* Here are some specific things to address in your report, in no particular order. This is *not* meant to be an exhaustive list.

- Did you perform any preprocessing on the data? If so, describe these steps (and cite sources if appropriate).
- What classification models did you build? How do they compare in terms of performance? What was the best performing model, and how did it do?
- Has your approach been used for this problem before²? If so, how do your numbers compare to the published results? If not, how do your results compare to the most similar/related approaches?
- What was your model-building and tuning regime? How did you address overfitting? How did you make hyperparameter choices (for example, type of regularization, strength of regularization, value of k in k -NN, distance metric in k -NN, etc.)?

¹Original source: <http://yann.lecun.com/exdb/mnist/>

²Answer: almost certainly yes.

Your primary deliverable for this assignment is a PDF report. The report should live in the highest-level directory of your GitHub repository. All of your code should also be present in the repository. The code should be easy to read and documented according to PEP-8 standards.

Add one last section to the report entitled “Contributions.” This section should highlight what each person contributed to the project, which should be supported by your GitHub commits.

Deadlines

Recommended Timetable

Here’s a recommendation for how to budget your time over the next couple of weeks as you work on this assignment.

- **Sep. 20–23:** Explore the dataset, think about feature engineering, build your first models.
- **Sep. 24–27:** Run more thorough experiments (hyperparameter tuning, further feature engineering, etc.), analyze your results and iterate, search the literature for related work on the problem, write your introductory materials for your report.
- **Sep. 28– Oct 1:** Complete experiments, take a step back and think about your report’s narrative, write draft of your *Experiments* and *Results* section.
- **Oct 1–4:** Wrap-up any pending experiments, write the *Conclusions* section and the abstract, revise and proof-read the entire report and submit it for peer review