ACTU PS5841 Data Science in Finance and Insurance - Autumn 2023 (Y. Wang)
Assignment - 8
Assigned 11/2/23, Due 11/10/23

Cross Validation, Regularization

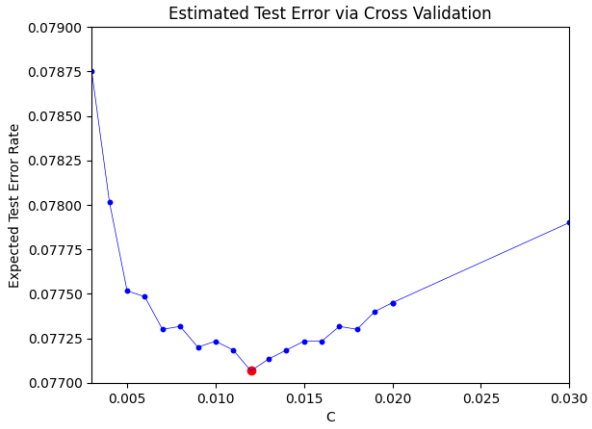
The file `mnist.npz` contains the MNIST dataset (which contains grayscale images of handwritten digits).

[*a*] Data wrangling to produce the standardized training set (60,000 images) and the standardized test set (10,000 images)

- Flatten each 28 × 28 image into a 1D array of 784 = 28 × 28 pixels.
- Standardize each of the 784 features into mean-zero and unit-variance ones by `sklearn.preprocessing.StandardScaler()`

[*b*] Fit a regularized logistic regression to the **standardized training set**, with the default values `C = 1.0` and `max_iter = 100`. Then evaluate the fitted model with the **standardized test set**, For this exercise, ignore the non-convergence warnings, if any.
Let your code report the training accuracy (0.94345) and test accuracy (0.9246).

[*c*] Code a 10-fold cross-validation to find the optimal C. Let your code report the optimal value of C and produce a plot similar to the following.



[*d*] Fit a regularized logistic regression to the **standardized training set**, with the optimal value of C found in [*c*] and `max_iter = 100`. Then evaluate the fitted model with the **standardized test set**.
Let your code report the training accuracy and test accuracy.