Problem Set 6

Data Science 602: Data Analysis and Machine Learning

Spring 2022

For the below problems, please use the `MNIST_784` data set from OpenML. Prior to using the data, scale the data and split into a test and training dataset. Use the first 60,000 images as training data, and the remaining 10,000 images as test data.

1. **Dimensionality Reduction** Using principal component analysis, reduce the dimensionality of the MNIST images to include 75% of the original variance. How many components remain following the dimensionality reduction?

2. **Support Vector Machines**. Use a support vector machine to classify whether a digit is less than 5 (i.e., $y \in \{0, 1, 2, 3, 4\}$). Find a set of hyperparameters, to include the kernel function and C, that maximize the F1 score.

   Notes:

   - As in problem set 5, you may want to initially search C over several orders of magnitude. Consider initially searching with `np.logspace` to search over orders of magnitude.
   - The hyperparameter selection may take a long time to run. If using Google Colab, you may want to save or print the model so the work is not lost if the model reconnects. See https://scikit-learn.org/stable/modules/model_persistence.html for details.
   - Use a random search (RandomizedSearchCV) to test a broad set of different hyperparameter values.

3. **Evaluation** Using the best estimator you found in the previous problem, show the confusion matrix for both the training and test data.