Dimensionality reduction: Applying PCA to the MNIST handwritten digits dataset

December 9, 2015

The MNIST training set is a collection of $60,000$ handwritten greyscale digits that have been centered and resized to $28x28$ pixels. It is one of the most popular data sets for testing machine learning algorithms. In this exercise we will learn to use the built-in PCA module in scikit-learn by applying it to parts of this data set.

Question 1: getting started

Use the provided mnist.py python module to load the MNIST data set and display a bunch of '1' digits using the **mnist.montage** function we have provided in that module.

Question 2: using PCA to reduce the dimension to 2

Use sklearn.decomposition.PCA to reduce the dimensionality of the '1' digits data set to 2 and draw the result as a scatterplot. Do the same for the '3' digits. Is there a qualitative difference between these scatterplots? Do you have any explanation why?

Question 3: interpreting the principal axes

Note that the principal axes are vectors of length 28×28, so they can be drawn as images, just like the digits themselves. Draw the mean and the first 5 principal axes '1' digits.
Can you describe in words what kind of variation do these principal axes capture in the data?
Hint: After calling **pca.fit** the sample mean is available at **pca.mean_** and the principal axes are stored at **pca.components_**.