ECE 407

HW 3

Feb 23, 2017

Handwritten digit recognition using a Gaussian generative model.

In class, we mentioned the MNIST data set of handwritten digits. You can obtain it from:

http://yann.lecun.com/exdb/mnist/index.html

In this problem, you will build a classifier for this data, by modeling each class as a multivariate (784-dimensional) Gaussian.

. (a) Upon downloading the data, you should have two training files (one with images, one with labels) and two test files. Unzip them. Load the data into MATLAB (you can use any other platform that you are familier with including Python)

. (b) Split the training set into two pieces – a training set of size 50000, and a separate validation set of size 10000. Also load in the test data.

. (c) Now fit a Gaussian generative model to the training data of 50000 points:

- Determine the class probabilities: what fraction π_0 of the training points are digit 0, for instance? Call these values $\pi_0, \ldots, \pi_9$.

- Fit a Gaussian to each digit, by finding the mean and the covariance of the corresponding data points. Let the Gaussian for the jth digit be $P_j = N(\mu_j, \Sigma_j)$.

  Using these two pieces of information, you can classify new images x using Bayes' rule: simply pick the digit j for which $\pi_j P_j(x)$ is largest.

. (d) One last step is needed: it is important to smooth the covariance matrices, and the usual way to do this is to add in cI, where c is some constant and I is the identity matrix. What value of c is right? Use the validation set to help you choose. That is, choose the value of c for which the resulting classifier makes the fewest mistakes on the validation set. What value of c did you get?

. (e) Turn in:

- All your code.

- Error rate on the MNIST test set.

- Out of the misclassified test digits, pick five at random and display them. For each instance, list the posterior probabilities Pr(y|x) of each of the ten classes.