MNIST with Naive Bayes and Logistic Regression

Timothy C. Havens

For this assignment, you will build Naive Bayes and Logistic Regression classifiers for the MNIST data set. Implement each part and present your code, results, and analysis as a Jupyter Notebook (submit both the notebook file and a PDF).

Part 1. (40 points) Build a Naive Bayes classifier for the the MNIST data set and test it on the test data set. Provide a confusion matrix, accuracy for each digit, and overall accuracy. Comment on your results.

Assume that the probability model for each pixel is a Gaussian distribution and that the prior probability of each class—i.e., digit—is equal. That is, $P(c=0) = P(c=1) = \ldots = P(c=9)$. Let $\mathbf{x} = (x_1, x_2, \ldots, x_{784})$ be the vector of pixel values for a given image and c is the class or digit, 0 to 9. Hence,

$$P(c|\mathbf{x}) = P(\mathbf{x}|c)P(c)/P(\mathbf{x}) \approx \prod_{i=1}^{784} P(x_i|c),$$

where $P(x_i|c) = \mathcal{N}(\mu_i, \sigma_i|c)$ is modeled as a Gaussian distribution from the training data. Train your Naive Bayes classifier and apply it to the test data set. For each of the 10 classes, show an image of the 784 conditional probabilities as a 28×28 image. That is, show an image of $P(x_i|c=1)$, $i = 1, 2, \ldots, 784$ as a 28×28 image, repeat for $c = 2$, etc. Do these images provide any insight into how this classifier works?

Part 2. (60 points) Build a regularized logistic regression classifier, where you use ridge (ℓ_2) regularization. Test this classifier on the MNIST data set by developing 10 classifiers: 0 versus all, 1 versus all, 2 versus all, ... , 9 versus all. Provide a confusion matrix, accuracy for each digit, and overall accuracy. Plot the overall test accuracy versus the regularization value where a log-scale is used for regularization value.

Essentially, the '1 versus all' classifier is trained to give you a probability of the digit 1 versus all other digits. Hence, digit 1 is class +1 and all other digits are class -1. Hence, to classify a test image, you take the maximum probability from all 10 classifiers, giving the predicted class of the input image.

ℓ_2 regularized logistic regression uses the following log-likelihood,

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{\lambda}{2}\|\mathbf{w}\|_2.$$

Recall, that by taking the gradient of $\mathcal{L}(\mathbf{w})$, we obtained the following gradient descent update equation for (non-regularized) logistic regression,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \mathcal{L}(\mathbf{w}),$$

where

$$\nabla \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \frac{-y_i \mathbf{x}_i \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}.$$

Derive the update equation for the regularized logistic regression and present your derivation. Apply that classifier to the MNIST data set. Don't forget to add a column of 1s to your 784-length feature vector (the image values) so that you get the bias term with your classifier. For each of your 10 trained classifiers, show an image of the 784 weights (don't show the bias weight) as a 28×28 image. Do these images provide any insight to how this classifier works?

Your score on each part of this assignment will be based both on the accuracy of the results and the presentation. Plots and images should be labeled, results should be described, and analysis should be provided that gives evidence for your claims. Don't just present a table or figure without describing the results in it. Tell the reader (or the grader as it may be) what they should glean from each table and figure.