Estimation of Mutual Information for Image Classification

Cosmo Harrigan

March 2019

Abstract

Entropy and mutual information are estimated on a discrete alphabet consisting of black and white images of the MNIST handwritten digit dataset. These information theoretic quantities are then used to perform supervised image classification, and the resulting performance is analyzed using single pixel features and top-k pixel features for prediction.

1 Dataset

We use a binarized version of the MNIST handwritten digit dataset. Each black and white image is represented as a binary vector of length 784, which is the produced by flattening the 28x28 pixel images. The class labels are the digits from 0 to 9. We use a training set of size 5000 and a separate test set of size 5000. An example of a binarized digit is shown in Figure 1. The code written to produce these results is available at:
`https://github.com/cosmoharrigan/entropy-estimation`.

2 Information Estimation

The MNIST domain is a supervised learning problem in which we are given a limited sample of labeled training data from which we can estimate the underlying distribution which generated the data. Based on this estimate, we can then classify new data. The objective is to maximize the accuracy achieved on the new data.

We estimate the entropy, conditional entropy, joint entropy and mutual information based on the counts in the training data without using smoothing. We note that smoothing may result in improved estimates [1].

The entropy of the class labels is calculated to be 3.318 bits, which is very close to the entropy $\log_2 10 = 3.322$ of a uniform distribution over digits.

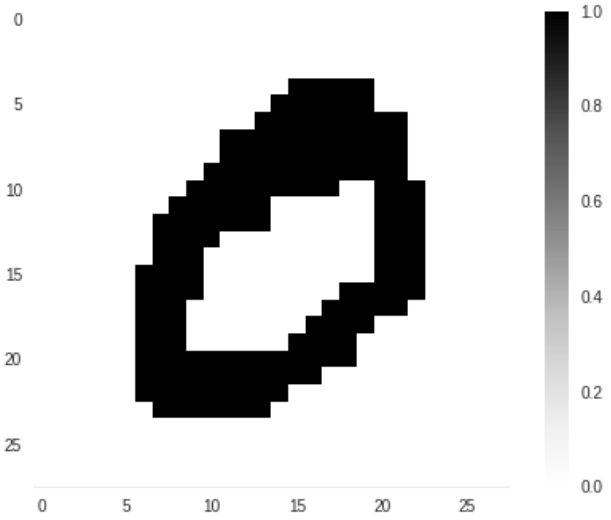

Figure 1: Example binarized digit image

3 Mutual information with individual pixels

We begin by examining the predictive capacity of individual pixels for predicting the class label. Estimating the mutual information $I(X;Y) = H(Y) - H(Y|X)$, we find that certain pixels have no mutual information with the class label, while the best performing pixels have mutual information of 0.349 bits. The mean mutual information is 0.091 bits. In Figure 2, a heatmap of the mutual information for each individual pixel with the class label is plotted on a 28x28 grid.

In Figure 3, the prediction accuracy for each individual pixel on the training set is illustrated for comparison. These results are obtained by determining the class label with the highest probability based on the number of times it occurs along with each value of a pixel. These assignments are then used to classify each data point. Comparing these two figures, it is evident that the mutual information of individual pixels with the class label, and their predictive accuracy on the training set, are closely related but not identical.

The maximum mutual information achieved using a single pixel was 0.349 bits, which is very far from the upper bound of 3.318 bits given by the entropy of the class labels. As a result, we will need to use more complex features to increase the predictive capacity.

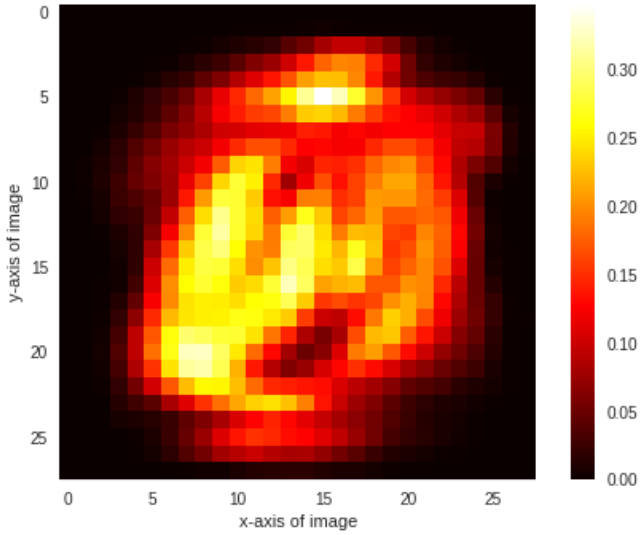

Figure 2: Mutual information of each pixel with the class label

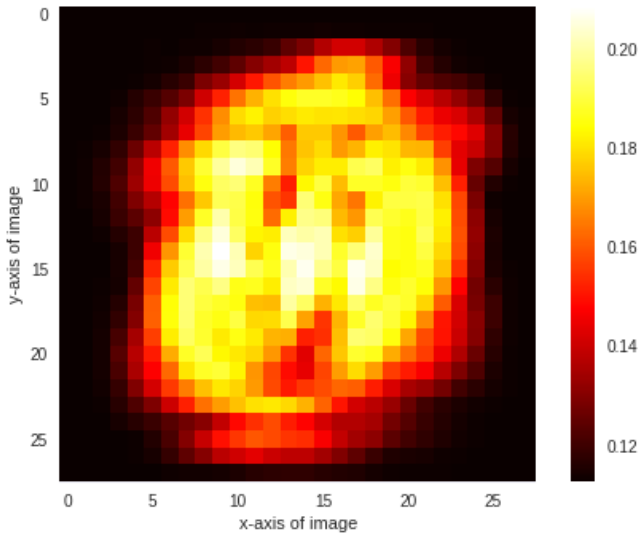

Figure 3: Prediction accuracy of each pixel on the training data

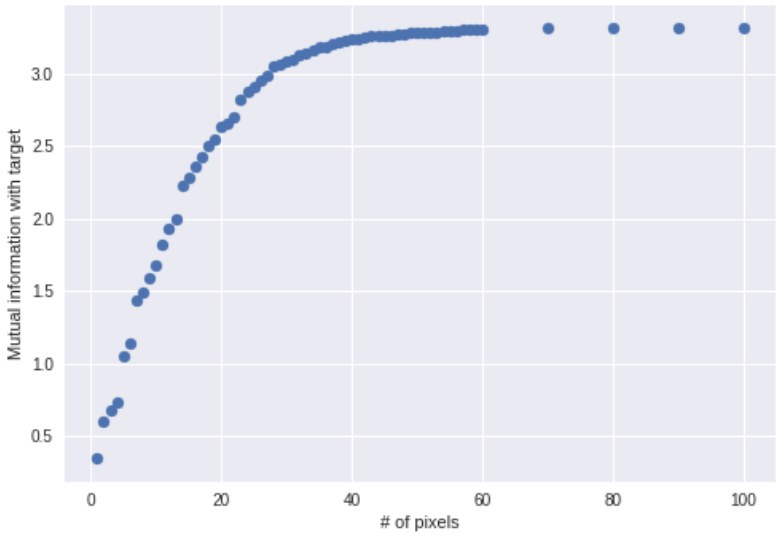

Figure 4: Mutual information of the top-k pixels with the class label

4 Mutual information with top-k pixels

In this experiment, we use the results from Section 3 to rank the pixels in descending order of their mutual information with the class label. We then use the top k pixels as our feature vector, and we estimate the mutual information between these pixels and the class label for a range of values of k. We evaluate the performance of each corresponding classifier on test data.

The results are illustrated in Figure 4. First, the top k pixels based on their individual mutual information with the class label were chosen to construct a feature vector. Then, the statistics of their joint distribution were used to estimate the mutual information between the feature vector and the class label. At first, the mutual information rapidly increases with the size of the feature vector, and then asymptotes near the entropy of 3.318 bits of the class label. At 30 pixels, the mutual information has already exceeded 3 bits.

5 Predictive accuracy on test set

The top-k mutual information pixels reached the maximum possible mutual information with their class label, implying that such a scheme may be able to be used to predict new data from the same underlying distribution. We evaluated the performance of the associated classifier on the test set, for a range of values of k.

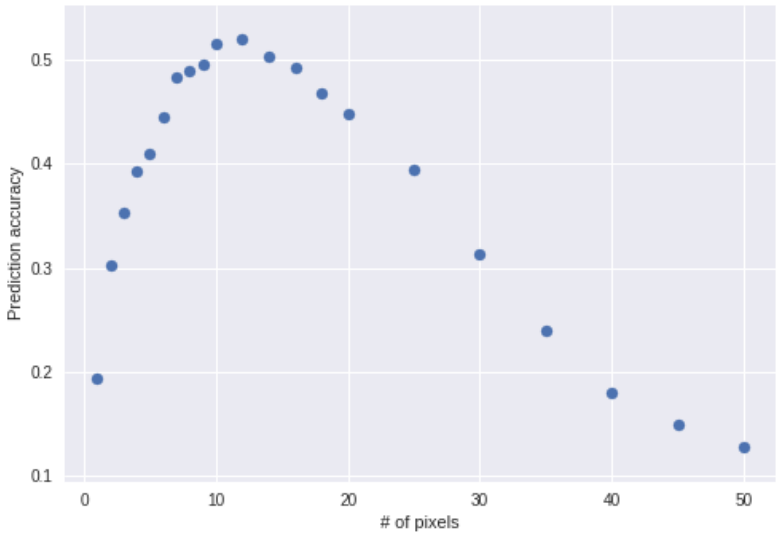

Figure 5: Prediction accuracy of the top-k pixels on the test data

In Figure 5, the prediction accuracy of the top-k pixels is illustrated. A striking feature of this plot is that the prediction accuracy initially increases with k, reaching a maximum at $k = 12$, and then decreasing again. The increase in prediction accuracy is significant, starting just under 20% accuracy using single pixels as features, and reaching a maximum of over 50% accuracy using a pixel vector of the top-12 pixels. As k is increased further, the accuracy rapidly drops off and eventually becomes worse than the single pixel accuracy.

6 Analysis

The fact that there is an intermediate value of k pixels which produces superior prediction accuracy suggests that the model may overfit to the training data if the length of the feature vector increases such that it is unlikely that sufficient samples were present in the training data to provide a representative sample for estimating the distribution.

The predictive accuracy of the top-k feature vector scheme is far below that of common classification algorithms. However, it is an extremely simple method which achieved initial performance of over 50% accuracy without any additional modifications. It is an illustration of the estimation of entropy and mutual information on discrete alphabets of images, and the application of these information theoretic quantities to feature selection for classification in supervised learning.

7 Discussion and future work

These basic information theoretic methods for feature selection for supervised classification provide a starting point for further analysis. I include several suggestions and references for alternative approaches below.

The *information bottleneck* method [2], based on rate-distortion theory [3], attempts to find a compact summary T of the data X while preserving relevant information about the predicted variable Y by minimizing the quantity

$$\min_{p(t|x)} I(X;T) - \beta I(T;Y)$$

Recently, an extension of this method, called the *deterministic information bottleneck*[4], has been proposed. This revised method replaces the $I(X;T)$ term, which is from channel coding, with an entropy term $H(T)$ from source coding. While the original method rewards stochasticity in the mapping from X to T, the revised method rewards a compact description of T and results in a deterministic encoding.

These methods provide an alternative approach for information theoretic clustering which could also be used as the basis of a classifier for supervised image classification.

The scheme for constructing feature vectors by combining a small subset of image pixels could also be extended to allow the composition of image patches as well as hierarchical combinations of image patches.

References

[1] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[4] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *Neural computation*, vol. 29, no. 6, pp. 1611–1630, 2017.