Abstract

We review a novel algorithm for unsupervised learning called the Information Sieve. We discuss its information theoretic background and the algorithm. Then we compare this model to another novel approach to learning called Generalized Low Rank Models. We discuss the qualitative appeal to both, and then compare results between the two models on 3 benchmark applications: Lossy Compression and Inpainting with MNIST digits, and classification using the MADELON dataset.

Background

For our project, we were interested in the algorithm development option. We were especially interested in learning about any new approaches to unsupervised learning, a hot topic in Machine Learning. Presently, there is an increased focus on learning deep representations of unlabeled data. This unsupervised learning context is where we find the popular artificial neural network for example. Hierarchical clustering is another example where we find multiple layers of information learned from our data. Methods like these involve the discovery of *latent* variables, which is essentially the meaningful information contained in the data. In our research, we found a new algorithm presented at ICML in 2015, (*reference "Sifiting Common info..") and further updated at ICML in 2016 with the additional paper "The Information Sieve". The Information Sieve (IS) was appealing to us due to its seemingly approachable nature. A large part of the appeal of the IS is the intuitive explanation of the methodology. In the author's own words:

> The information sieve introduces a new way of learning things piece by piece… We pass the data through the first layer of the sieve to extract the "most informative" pattern in the data, the data is transformed and the remaining information trickles down to the next layer of the sieve. This "remainder information" contains all the information from the original data except for what was already learned. This allows incremental learning that is guaranteed to improve at each step, and to never duplicate effort by re-learning what is already known.

The author claims that, as opposed to many other models where signal is learned from the data "all at once," the Information Sieve algorithm is more reminiscent of human learning. (https://apparenthorizons.com/2015/07/20/the-information-sieve-with-bonus-eigen-faces/).

In other words, in each iteration of the IS algorithm, data, X, is "passed through the sieve" to extract a single latent factor, y. This factor is constructed so as to minimize the "Total Correlation" in the data when conditioned on y. Once the factor is extracted, the data is transformed to X' to remove its dependence on y. Then the transformed data can be passed through the next layer of the sieve and the process repeats until all common information has been extracted and the remainder contains only independent noise.

Mathematical Concepts behind The Information Sieve

IN Midterm Report

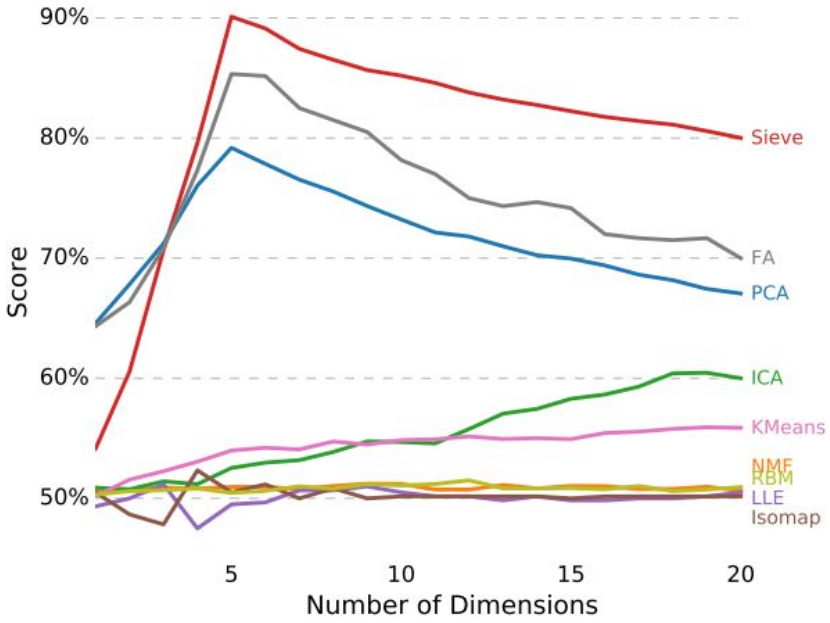
Problem

We aim to compare the Information Sieve against Generalized Low Rank Models at the three primary tasks of: Lossy Compression, Imputing (inpainting), and Classification.

In lossy compression, as the name suggests, the aim is to store the target information in a smaller amount of memory than the original. In this case, it is not necessary to maintain the fidelity of the file. Compression ratio is prioritized over accuracy so long as the original information may be understood. A common use is the lossy compression of images where a loss in quality isn't necessarily detrimental to the image. Here we directly compare both methods at compressing 60000 digits in the MNIST dataset.

The second task is inpainting which is a subset of the problem of imputing data. The bottom half of images will be removed and must be reconstructed using only knowledge of other complete images and the top half of the incomplete images. Because there is quantitative benchmark of success provided in the sieve paper, a qualitative visual approach will be taken to determine success. Again, we will compare both methods on inpainting the digits in the MNIST dataset.

Finally, we will compare both methods on a difficult classification problem using the MADELON dataset. In the paper where the Information Sieve is first introduced, "Sifting Common Information from Many Variables" (Ver Steeg), the IS was evaluated against some standard models on a couple of classification tasks. One, where the IS performed better than all other evaluated competitors was using the MADELON dataset.

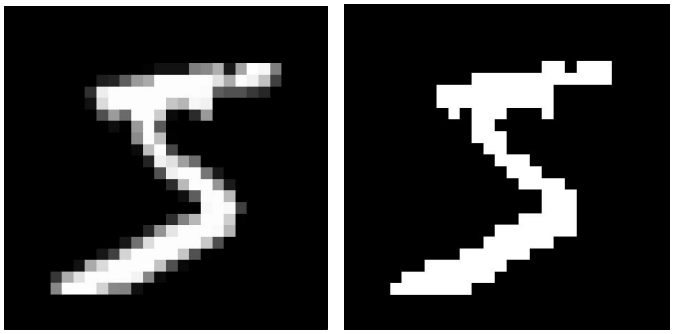


We see that beyond about rank 3, the IS outperforms all of the methods to which it was prepared. The closest competitor being "factor analysis". However, GLRMs, being a new method itself, were not evaluated again the IS.

Dataset

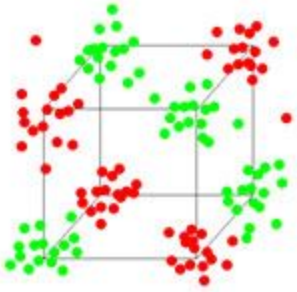
MNIST

The MNIST database consists of 70,000 examples of handwritten digits centered in a 28x28 pixel image. These images are presented in a compressed file format that, when extracted into a csv file, yields rows that represent individual images of digits. The first column is an integer from 0 to 9 representing the handwritten digit encoded in the remaining 784 columns which represent grayscale values from 0 to 255 of the pixels of the image. After downloading, we process this data identically as described in the sieve paper by first normalizing the values and then rounding them at a threshold of 0.5 so that the image is effectively composed of black and white pixels. Of these 70,000 examples, the database is split into 60,000 training examples and a test set of 10,000 samples. This dataset was made available to download directly in csv format for both the training and test datasets.



Madelon

The MADELON dataset is an artificially created binary classification dataset designed to have many confounding factors. The data is generated by first creating 32 independent Gaussian clusters and introducing some covariance between them. These clusters are then placed at random on the vertices of a five dimensional hypercube and randomly labeled either +1 or -1. The five dimensions represent the 5 informative features to which 15 linear combinations of them are added to form additional redundant features. Finally, 480 additional features were added with no predictive power for a total of 500 features. This dataset was also made available in downloads of the training and test datasets but with separate downloads for the classification labels.



Methodology

GLRM for Lossy Compression of MNIST Digits

The first comparison we ran against the Information Sieve was the task of lossy compression on the MNIST dataset using GLRM. The sieve claims a reduction of file size of about 70% when compressing 50,000 digits using 100 sieve representatives. This brings the bits used to store an MNIST digit from 784 to 243. Because the performance of GLRM compression depends on the rank of the model, a rank had to be chosen that would produce a similar compression rate per MNIST digit so that the visual quality of each image can be directly compared.

GLRM finds a low rank factorized representation of the observed data Y. Thus our compressed representation of our images is stored in two matrices X and W whose product is approximately our original data Y. Matrices X and Y are real valued and each entry requires 64 bits of memory to hold. If we use approximate values, this memory requirement can be reduced to 16 bits. The total memory required is now the sum of the number of entries in X and W multiplied by 16 bits. Matrix X is of size n*k where n is the number of compressed MNIST digits and k is the rank of the model. Matrix W is of size k*784 where 784 is the number of bits in the original image (each bit represents a feature in the dataset). Having established this, we can calculate a formula for the approximate file size of storing n MNIST digits using GLRM of rank k : (16n + 12544)k where the cost of storing an additional digit is then 16k.

One interesting feature arises from this calculation: there is an initial fixed cost to storing information using GLRM. This indicates that it is a method best used for storing large amounts of data. Now, considering the compression rate of the sieve, we may calculate that using a model of rank 15 will yield a similar average compression rate to the sieve of approximately 244 bits per digit.

Overall, the author used 50,000 digits of the MNIST data to perform this lossy compression. Our compression utilized all 60,000 training digits however, this discrepancy would only serve in the sieve's favor because more observations are fitted within the same rank model. The author also points out that spatial information was not utilized and emphasizes this point by applying a consistent random shuffling of the pixels of each digit. We held ourselves to the same constraint in fitting our GLRM because glrm does not take into account spatial proximity of features.

InPainting

Another Application of the Information Sieve presented by Ver Steeg is "In-Painting". Using MNIST digits once again, the author removed the bottom half of the image for a subset of the digits and then attempted to fill them back in using the latent variables derived from the algorithm. With the Information Sieve, "missing data is handled quite gracefully" as you are simply able to optimize results over the observed values only.

Due to both capacity constraints in Juliabox and time constraints in general. GLRMs were only trained on

GLRMs are also well designed for handling missing data. With GLRMs we are able to impute unobserved values. We removed the bottom portion of the image for 10%

of our digits and fit a rank (k) model to the data. Due to limitations in computing power, we did not use 50,000 training digits as Ver Steeg did. We used only (n) digits.

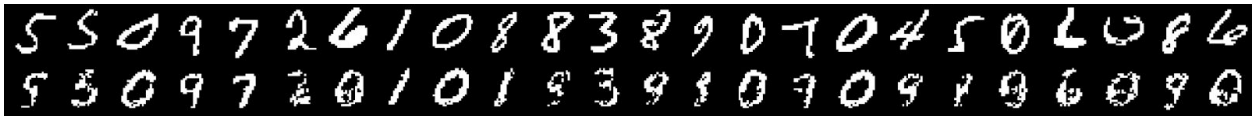
Classification

Finally,

Comparison

Lossy Compression

Ultimately, GLRMs performed as well, if not better, than the Information Sieve at the same compression.



Inpainting

Shit

Madelon

Shit

Conclusion

Overall, the Information Sieve dramatically outperformed our implementations of Generalized Low Rank Models.

What we do note, however is that the sieve is a particularly versatile tool unaffected by validation and scaling

Improvements
Inpainting
Madelon Classification

Bibliography

https://arxiv.org/abs/1606.02307

https://arxiv.org/abs/1507.02284

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

http://makeyourownneuralnetwork.blogspot.com/2015/03/the-mnist-dataset-of-handwritten-digits.html