

Introduction

In this assignment I will examine how well different clustering and dimension reduction algorithms compare. I will reuse one of my datasets from assignment one, that I will briefly review to see how well clustering algorithms classify each instance. I will also see how my best neural network learner from assignment 1 performs after projecting the dataset to a lower dimension. Sci-kit was used to run all algorithms.

Datasets

Dataset	Description	Characteristics
MINST	Handwritten digit images written by 500 writers.	Highly dimensional sparse dataset sampled to 10,000 instances and 3600 attributes
Balance Scale Data Set *used in Assignment 1	In balance scale experiments, a child is asked to predict the outcome of placing certain numbers of equal weights at various distances to the left or right of a fulcrum [Shultz]	The Balance Sale dataset to only had 4 variables and 650 attributes and was generated to model experiments conducted by Siegler, R. S

Why is my data interesting?

These MINST data set is unlabeled and will show the clustering algorithms handle sparse data with a high dimension. This dataset has 3600 dimensions, so after projection the data down to a two dimensional space it will be interesting see how much information we lose or gain about this dataset. Given the setting it will be difficult to model the data even with a single Gaussian. . As for the balance scale dataset it is a labeled dataset, so I can check to see if each instance of the same class is being clustered in similar clusters. This dataset is a multi-labeled dataset with 3 classes, and one of the labels only represents 7.84% of the data which could lead to new instances being mapped to the wrong class.

Algorithms

Expectation Maximization

Given k different mixtures of normal distributions we are unable to observe which instances were generated from which distribution. Therefore, we must introduce try to estimate the hidden variables associated for each observed variable. There are k hidden variables for each observed value, so each we can represent the full description of each datum as $\langle x_i, z_1, z_2, z_3, \dots, z_k \rangle$. The Expectation Maximization uses two algorithms to search for a maximum likelihood hypothesis by repeatedly re-estimating the expected values of the hidden variables z_{ij} given its current hypothesis, then recalculating the maximum likelihood hypothesis using the expected values for the hidden variables. Expectation, the first step is used to calculate the expected values of the hidden variables, this is then passed to the Maximization algorithm which tries to computer a new maximum likelihood hypothesis [6].

K Means

The goal of K Means is to divide M data points of N dimensions into K clusters based on a distance metric, so that the within-cluster sum of squared residuals is minimized [5]. The algorithm iteratively assigns instances to one of the K groups based on the values of its N features. In doing this we generalize our dataset which means that we are losing details on specific instances. K means is an extension of the EM algorithm.

Independent Component Analysis

Consider the fact that there is some data X of \mathbb{R}^n that has n independent sources [Ng].

Independent Component Analysis is an algorithm developed to transform a highly dimensional

multivariate data to a linear representation of non-gaussian data, so that the components are statistically independent [8].

Principal Component Analysis

Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the important information from the data table and to express this information as a set of new orthogonal variables called principal components [1].

Randomized Projections

The following dimensionality reduction lemma applies to arbitrary mixtures of Gaussians. Its statement refers to the notion of separation introduced earlier, and implies that by random projection, data from a mixture of k Gaussians can be mapped into a subspace of dimension just $O(\log k)$. Since it is conceptually much easier to design algorithms for spherical clusters than ellipsoidal ones, this feature of random projection can be expected to simplify the learning of the projected mixture [3].

Factor Analysis

Factor analysis operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable, which is known as reducing dimensionality.

Evaluation Metrics for Clustering Algorithms

Below are the evaluation metrics I decided to use for the clustering algorithms listed above.

Silhouette

In the MINST dataset since the dataset is unlabeled evaluation must be performed using the model itself. The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores. The mean distance between a sample and all other points in the same class. The mean distance between a sample and all other points in the next nearest cluster [7].

Calinski Harabaz

Calinski Harabaz index is a method for identifying clusters of points in a multidimensional Euclidean space [10]. It evaluates the cluster validity based on the average between- and within cluster sum of squares.

Akaike Information Criterion (AIC)

AIC is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

Homogeneity Score

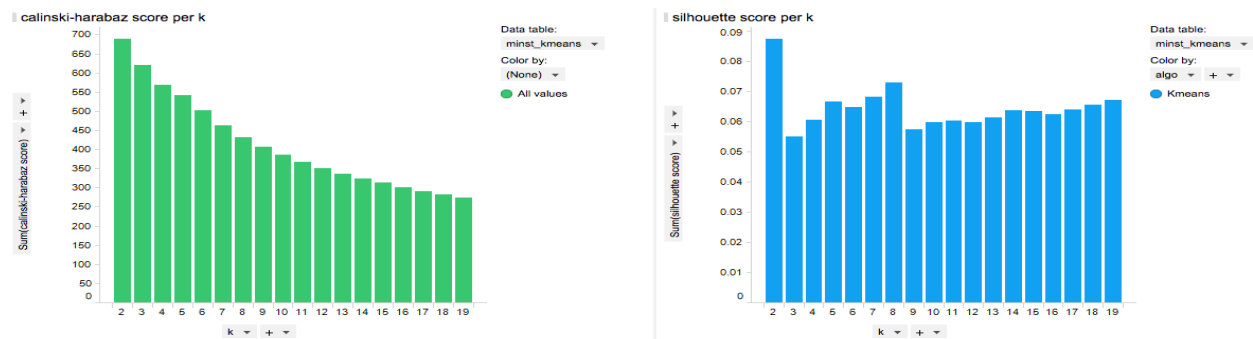
Given the knowledge of the ground truth class assignments of the samples, it is possible to define some intuitive metric using conditional entropy analysis. Homogeneity score measures how much each cluster containing only members of a single class [12] .

Analysis

MINST

Choosing K

Choosing k was different giving the dataset since MINST was unlabeled and the balance was labeled. The metrics I used for MINST were the silhouette and calinski harabaz score. Both let me know the within cluster dispersions. I Initially, chose k=8 because I felt as though this model had the best silhouette and calinski harabaz together. The silhouette score allowed me to observe the mean distance between a sample and points in the next nearest cluster. This is important because there could be overlap between clusters and shows me if each instance is labeled correctly or not. When clusters are dense and well separated the calinski harabaz was higher which is illustrated well in the diagram below to the left.



The figure above shows the calinski harabaz (left) and silhouette score (right) vs k values.

When I looked at the AIC chart for each model I saw that my initial guess for the k=8 value was supported and that the model I selected was a was a strong model. AIC let me know which model will minimize information loss which is great to use in the dimension reduction setting.

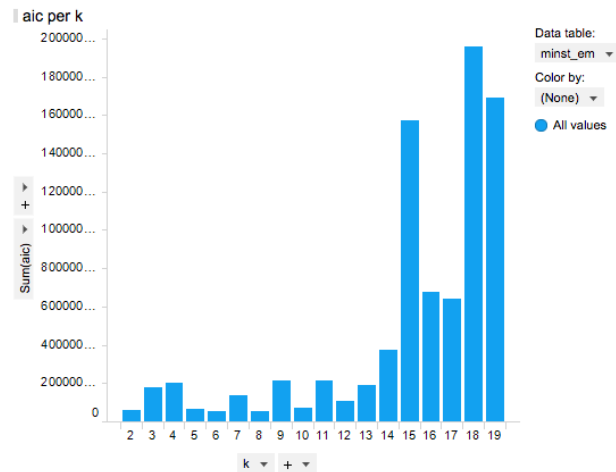
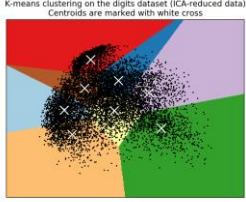



Figure above shows different MINST dataset models AIC score.

Dimension Reduction Clusters

Reduction Algorithm	Vis
Randomized Projection	<p><-means clustering on the digits dataset (Randomized Projection-reduced data) Centroids are marked with white cross</p>
Factor Analysis	<p>K-means clustering on the digits dataset (Factor Analysis-reduced data) Centroids are marked with white cross</p>

ICA	 <p>K-means clustering on the digits dataset (ICA-reduced data) Centroids are marked with white cross</p>
PCA	 <p>K-means clustering on the digits dataset (PCA-reduced data) Centroids are marked with white cross</p>

Balance Scale

Choosing K

Since this dataset was labeled, a good metric for cluster evaluation is the homogeneity score. As can be seen below after running all of the dimension reduction algorithms on the dataset and projecting the data to a 2 dimensional space. I decided to go with $k=6$ as my best model for this dataset.

Reduction Algorithm Comparison

Factor Analysis does a great job at detecting structure in the relationship between variables for classification. Factor Analysis gave a score of 1 at every iteration of k between 6 and 19 which led to my decision of choosing my best model to be $k=6$. Randomized projection on the other hand performs poorly partly due to the fact that this is a low dimensional dataset being projected to an even lower dimensional space. The other dimension reduction algorithms gave an homogeneity score around 50% meaning there is a 50-50 chance the algorithms are being clustered inappropriately. The green line just above that is the without any reduction algorithm performed on it, as seen below all, but randomized projection performed better than it.

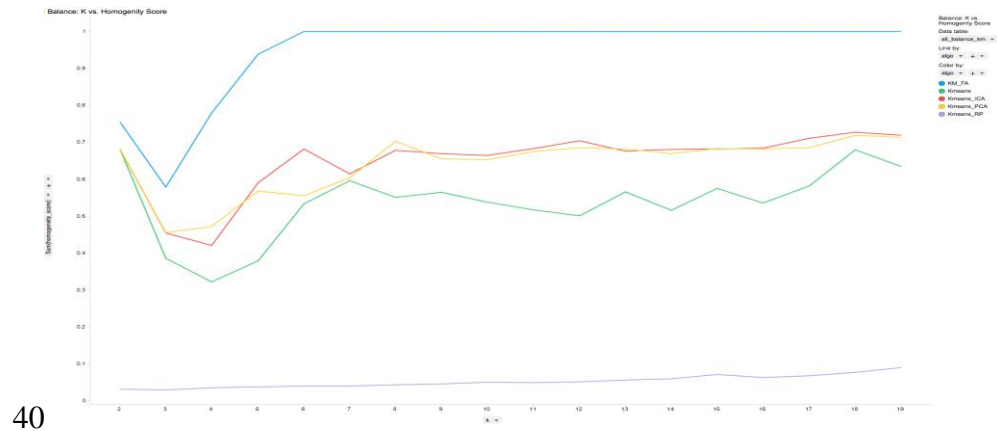
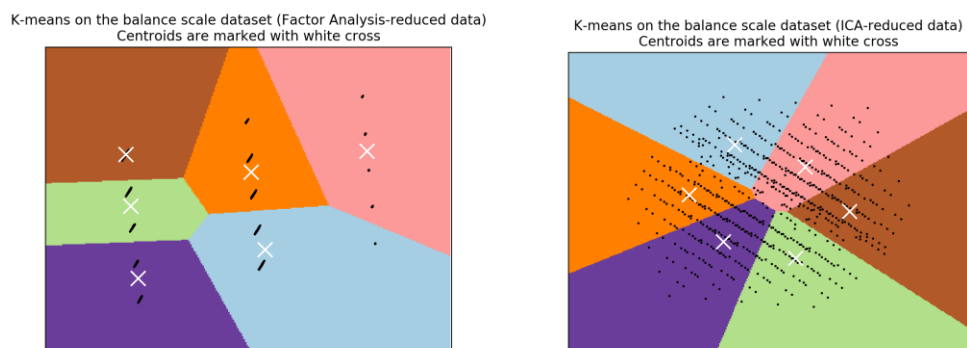


Figure above shows different Homogeneity scores for the balance dataset. The blue line is Factor Analysis, Red is ICA, Yellow is PCA, Purple is RP, and Green is regular Kmeans. (Legend is to the right)



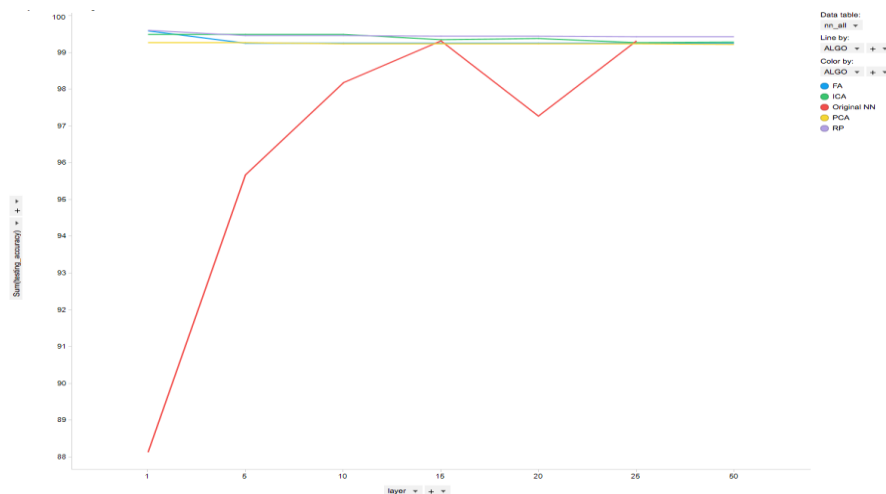
The figure above shows how factor analysis (left) and ica (right) clusters each data point.

Detailed Factor Analysis Observation

Because of a perfect homogeneity score after running Factor Analysis as my reduction algorithm I decided to dive deeper. As you can see above on the left Factor Analysis created a very tight

intra-cluster distances between instances after reducing the data, this supports the perfect homogeneity score. There also appears to be clusters within clusters which shows a diversity within each. The clusters are also very distinct, by this I mean they take on their own shape opposed to PCA's as you can see to the right. PCA's homogeneity score fluctuates between 40% - 60% , so the picture above is just a dispersion of instances and by chance they are landing in the right clusters.

Neural Network



After running my best neural network model from assignment 1 I observed that as the number of layers increased all of the dimensionality reduction algorithms performed better on my dataset. As you can see above randomized projection had the highest training accuracy as the number of layers increased. Overall Neural Networks perform well on this dataset, but after applying the reduction algorithms the performance is almost perfect. The reduction algorithms are doing a good job at finding out where each mixture model the latent variables belong to. In assignment 1 a support vector machine with a polynomial kernel was my best model emphasizing that there

are a lot of unobserved characteristics of this dataset and to find them some sort of transformation needs to occur.

Sources

- [1] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010): 433-459.
- [2] Dasgupta, Sanjoy. "Experiments with random projection." *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000.
- [3] Fern, Xiaoli Zhang, and Carla E. Brodley. "Random projection for high dimensional data clustering: A cluster ensemble approach." *ICML*. Vol. 3. 2003.
- [4] Hopcroft, John, and Ravi Kannan. "Computer science theory for the information age." (2012).
- [5] Hartigan, J. A., and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, 1979, pp. 100–108., www.jstor.org/stable/2346830.
- [6] Michalski, Ryszard Stanislaw, Jaime Guillermo. Carbonell, and Tom M. Mitchell. *Machine Learning*. Los Altos: M. Kaufmann, 1986. Print.
- [7] Rousseeuw, Peter J. , Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, Volume 20, 1987, Pages 53-65, ISSN 0377-0427, [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7).
- Shultz, Thomas R., Denis Mareschal, and William C. Schmidt. "Modeling cognitive development on balance scale phenomena." *Machine Learning* 16.1-2 (1994): 57-86..
- [8] Stone, James V. "Independent Components Analysis." *Wiley StatsRef: Statistics Reference Online* (2014): n. pag. Web.
- [9] Schwarz, Gideon. Estimating the Dimension of a Model. *Ann. Statist.* 6 (1978), no. 2, 461--464. doi:10.1214/aos/1176344136. <http://projecteuclid.org/euclid.aos/1176344136>.
- [10] T. Calinski and J. Harabasz, 1974. "A dendrite method for cluster analysis". *Communications in Statistics*

[11] Yong, An Gie, and Sean Pearce. "A beginner's guide to factor analysis: Focusing on exploratory factor analysis." *Tutorials in Quantitative Methods for Psychology* 9.2 (2013): 79-94.

[12] <http://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness-and-v-measure>