ISyE 6416 – Basic Statistical Methods – Spring 2016
Bonus Project: "Big" Data
Report

Team Member Names: Caroline Roeger, Damon Frezza

Project Title: Clustering and Classification of Handwritten Digits

Responsibilities:
- Caroline Roeger: Clustering
- Damon Frezza: Classification

Problem Statement

We tried to identify the best algorithm or method for handwritten digit recognition. Handwritten digit recognition is important in a world where tasks and processes are becoming more automated. Banking, for example, is significantly more automated than it used to be. In order to deposit checks, a person may insert a check into an ATM and the ATM will read the amount and account numbers. A person may also simply take a picture of their check with their phone and use a banking application to deposit the check. Both of these methods require precise handwritten digit recognition. Misclassification during these processes would be costly to either the recipient or the patron. Due to this we plan to compare several methods of both clustering and classification algorithms in order to see which method performs best.

Performance is measured in several ways. First and foremost, we are interested in the clustering/classification quality of the algorithms. This measure of performance will have the highest weight in our evaluation. The second measure of performance is speed or efficiency of the algorithm. The last measure is ease of implementation.

Data Source

The "Semeion Handwritten Digit Data Set", created by the Italian company Tattile and donated to the Semeion Research Center of Sciences of Communication (Rome), contains 1593 handwritten digits from about 80 persons (https://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit). Each person was instructed to write on a paper all the digits from 0 to 9 twice, once in a normal, careful way and once in a fast way. The images were scanned using a gray scale of 256 values, fitted into 16x16 pixel rectangular boxes and then each pixel was assigned a Boolean value (0 if the original value on the gray scale was below 127, 1 otherwise). The result is a data set of 1593 instances each having 256 binary attributes and a label indicating the digit. The data set has no missing values. Figure 1 displays five instances from the data set.

Figure 1: Digits from the Semeion Data Set

Clustering (Unsupervised Learning)

Clustering is the task of grouping a given set of data points into clusters in such a way that data points within a cluster are similar whereas points from different clusters are dissimilar. In contrast to classification there are no labels, i.e. clustering is an example of unsupervised learning. In general one can distinguish between methods where the number of clusters is not known a priori and methods where the number of clusters is given. Here we want to assign the 1593 digits of the Semeion data set to ten clusters. Thus, only the latter was considered in this project. Following algorithms were implemented and compared:

- **k-means clustering algorithm:** Starting with ten cluster centroids drawn randomly from the data set each data point is assigned to the closest centroid in terms of the squared ℓ_2 norm. The centroids are then updated as the mean of the data points assigned to it and the data points are reassigned to the now closest centroid. This is repeated until convergence of the sum of squared errors. The result is a hard clustering.
- **k-median clustering algorithm:** Starting with ten cluster centroids drawn randomly from the data set each data point is assigned to the closest centroid in terms of the ℓ_1 norm. The centroids are then updated as the median of the data points assigned to it and the data points are reassigned to the now closest centroid. This is repeated until convergence of the sum of absolute errors. The result is a hard clustering.
- **Fuzzy k-means clustering algorithm:** Fuzzy k-means clustering is similar to k-means clustering only that instead of assigning each data point to exactly one centroid, every data point has a degree of belonging to each of the centroids and this degree is related inversely to the distance between the data point and the centroid:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\|x_i - c_j\|_2}{\|x_i - c_l\|_2} \right)^{\frac{2}{m-1}}}$$

  where $m > 1$ controls how fuzzy the boundaries between clusters are (here values close to 1 gave the best results). The centroid is then updated as the mean of all points weighted by their degree of belonging to this centroid:

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m x_i}{\sum_{i=1}^{N} \mu_{ij}^m}$$

  This is repeated until convergence of the sum of weighted squared errors

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{ij}^m \|x_i - c_j\|_2^2$$

  The result is a soft clustering. (Bezdek, 2013)
- **EM clustering algorithm for a principle components Gaussian mixture model:** This algorithm combines the basic EM Algorithm for a Gaussian Mixture Model with a principle component dimension reduction applied to the covariance matrix to obtain less noisy estimates. In the E-step of the algorithm one computes the class membership distribution conditional on the current parameters and the data. In the M-step one then updates the parameter estimates given the current class membership distribution. To get less noisy estimates a spectral decomposition is applied to the covariance matrix resulting in a "rank-q plus noise"-estimate. This is repeated until convergence of the data log-likelihood. The result is a soft clustering. Two different initialization approaches are used. In the first case the EM algorithm starts from scratch with naïve initial parameters (means drawn randomly from the data set, identity matrices as covariance matrices, same likelihood for every cluster). In the second case the algorithm is applied on top of a k-means clustering,

using the k-means cluster assignments as initialization for the class membership distribution.

- **Non-negative matrix factorization clustering algorithm:** Using non-negative matrix factorization one approximates the data matrix X $(n \times m)$ as the product of a matrix A $(n \times k)$ and a matrix S $(k \times m)$, where n is the number of instances, m is the number of attributes and k is the number of clusters. The update rules for A and S are such that at convergence matrix A will indicate the clustering:

$$A \leftarrow A \cdot \frac{XS^T}{ASS^T}, S \leftarrow S \cdot \frac{A^T X}{A^T AS}$$

  The result is a soft clustering. (e.g., Tjoa and Liu 2010)

All algorithms were implemented in MATLAB from scratch and were compared with respect to clustering quality, run time as well as ease of implementation. To evaluate clustering quality following criteria were used:

- Visual inspection of cluster means: do the cluster means resemble the different digits?
- Purity: Each cluster is assigned to the class which is most frequent in the cluster. Then the purity is computed as the fraction of correctly assigned instances. In general a higher value is better. (Manning et al., 2008)
- Miscategoriaztion rate: Each class/label is assigned to the cluster which is most frequent for the class. Then the miscategorization rate is computed as the fraction of incorrectly assigned instances. In general, a lower value is better.
- Sum of squared errors: Sum of squared errors between the samples and their corresponding centroids (weighted distances in case of soft clustering). In general, a lower value is better.

It should be noted that miscategorization and purity are complementary: purity can be optimized by assigning every instance to a different cluster whereas miscategorization can be minimized by assigning all instances to the same cluster. Thus, an effective clustering algorithm should achieve good results for both measures.

Classification (Supervised Learning)

Classification is the task of properly classifying a new observation based on the model parameters tuned using a training data set consisting of data for which the classes are known. This is a type of supervised learning because we know the classes for all the data in the training set and we will use those labels to tune the model. In order to compare the supervised methods, we will randomly split the data into a training set and a testing set. After training the models, we will analyze and compare the performances on the test set. Following algorithms shall be compared:

- **Linear Discriminant Analysis (LDA):** Two starting assumptions are that p(x|y) is normal and that the variance covariance matrices for each class are equal. The equal variance assumption is what leads to the linearity of LDA. After estimating the mean of each class and the overall variance, the discriminant functions are calculated. Using these, new observations are assigned to the class for which the value of its discriminant is the highest. (Nongpiur et al., 2013)
- **Multinomial Regression (softmax):** This generalization of logistic regression is used to predict probabilities of different possible outcomes for multi-class systems. The softmax function is used to act as a probabilistic indicator function which is conveniently differentiable, returning 0 for values less than the maximum of all values. The softmax function is the classification criteria for the multinomial method.

    - Lasso variant: ℓ_1 regularization is a promising variant for handwritten digit recognition because the data are sparse. Many pixels on the image are not utilized and are relatively unimportant when trying to classify new observations. ℓ_1 regularization performs model selection and eliminates unnecessary predictors from consideration for classification. (Koh et al., 2007)
- **Naïve Bayes:** This assumes that the predictors are independent conditional on the response category. The characteristic equation which describes the method is $p(y|\boldsymbol{x}) \propto p(y)\prod_{j=1} p(x_j|y)$. The one dimensional conditional distributions are generally much easier to estimate than the joint conditional distribution. The reduction in dimensionality drastically reduces complexity but the method hinges on a strong assumption. The Bayes classifier minimizes the expected loss conditional on the observed data.
- **Support Vector Machine (SVM):** An SVM model is a representation of the observations as points mapped to a new space in higher dimensions so that the categories can be separated by a clear gap that is as wide as possible. New observations are mapped into the same space and are classified based on their position in that space. SVM is usually only used with two classes. In order to generalize SVM to a multiclass scenario, we plan to use the "one vs one" design with K(K-1)/2 binary SVM models (K=10 in our case). Hsu and Lin (2002) demonstrate that the "one vs one" design outperforms the "one vs all" on several data sets and the differences are greater for larger data sets.
- **Neural Networks:** Artificial neural networks are a class of multi-layer hierarchical variable models which were inspired by the architecture of the brain. The inputs are transformed recursively into layers of hidden variables which are then transformed into response predictions. The data transformations at the inner layers are generally the composition of a linear transformation and a non-linear scalar function. Backpropogation is used as a gradient descent algorithm selecting the best parameters at each layer of the network. New observations are transformed by each layer and are classified based on their transformations.

Results

a) Clustering

Table 1 summarizes the results for the different clustering algorithms where the quality measures are computed as averages and medians over ten independent iterations. Figure 2 and Figure 3 further display exemplary means resulting from an effective clustering algorithm (Fuzzy k-means, Figure 2) and a poorly performing clustering method (Non-negative matrix factorization, Figure 3). As one can see there are significant differences in quality, run time and ease of implementation. It should be noted, though, that the clustering quality as well as the run time greatly depend on the initialization and for many clustering algorithms there exist sophisticated initialization methods (e.g., Blömer and Bujna, 2013). However, in these cases great parts of the clustering are actually already performed by the initialization making a fair comparison difficult. Therefore, even though we tested different initialization methods, for comparisons all algorithms except for "EM with k-means initialization" were initialized randomly as described in the previous section. Based on our results fuzzy k-means and EM with k-means initialization can be considered the most effective clustering methods for the handwritten digit recognition task. On the other hand, EM and non-negative matrix factorization did not provide good results when used with naïve initializations. Nonetheless, as one can see from Figure 2 even the better methods were not able to reliably distinguish all ten digits (especially the number 5 seems to be hard to identify). Thus, there is still room for improvement.

	k-means	k-median	fuzzy k-means	EM – random initialization	EM – k-means initialization	Non-negative matrix factorization
Miscateg. Rate (mean/median)	0.3927/ 0.3914	0.4599/ 0.4545	0.3889/ 0.3873	0.4901/ 0.4765	0.3652/ 0.3685	0.5289/ 0.5348
Purity (mean/median)	0.5929/ 0.5920	0.5412/ 0.5414	0.6045/ 0.5979	0.4982/ 0.5135	0.6211/ 0.6070	0.4307/ 0.4284
SSE (mean/median)	6.6845e+04/ 6.6831e+04	6.7800e+04/ 6.7738e+04	6.6618e+04/ 6.6580e+04	7.0881e+04/ 7.0664e+04	6.7092e+04/ 6.7082e+04	7.9940e+04/ 7.9904e+04
Visual Inspection	**	*	**	*	**	*
Overall Quality	**	*	***	*	***	*
Run Time	**	**	**	*	**	**
Ease of Implementation	***	***	***	**	**	***
Ranking	3	4	1	5	2	6

Table 1: Results Clustering (*** = good, ** = ok, * = poor)

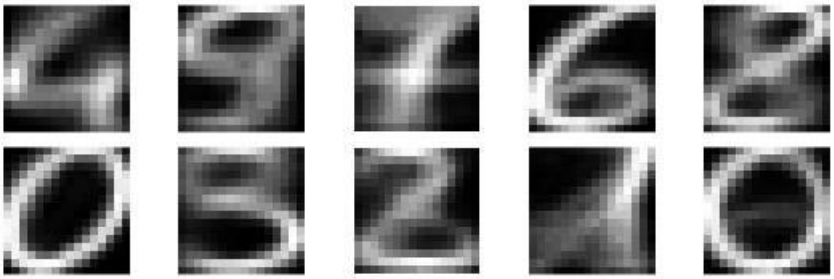

Figure 2: Means Fuzzy k-means

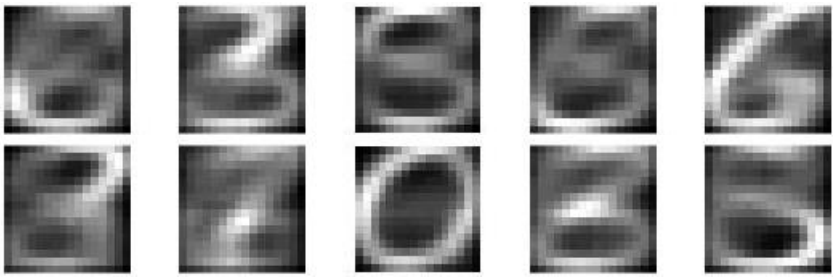

Figure 3: Means Non-negative Matrix Factorization

b) Classification

Table 2 summarizes the results for the different classification algorithms where the quality measures are computed as averages and medians over twenty iterations of 10-fold cross-validation. We include measures to elicit which digits are the easiest to classify and which are the hardest. 7 appears to be most difficult along with 4. 0 and 8 appear to be the easiest to properly classify. SVM outperforms all the other classification methods consistently and is the most reliable method we study. The Neural Network method also performs quite well but was disproportionately difficult to implement. The Lasso multinomial regression technique performed similarly to the Neural Network but was significantly easier to implement so we rank it better. Conversely, multinomial regression without the Lasso model selection performed the poorest of all classification methods. The model selection reduced the influence of noise and variability based on different people's handwriting. SVM has a 6% misclassification rate, which may still be too large to be reliable for banking transactions.

	Linear Discriminant Analysis	Multinomial Regression	Lasso Multinomial Regression	Naïve Bayes	Support Vector Machine	Neural Network
Misclass. Rate (mean/median)	.1066/ .1132	.2591/ .2704	.0843/ .0852	.1516/ .1541	.0660/ .0629	.0865/ .0881
Max Misclass. rate(digit)	.0570 (7)	.0380 (7,9)	.0127 (7)	.2298 (4)	0 (ALL)	.0186 (4)
Min Misclass. rate(digit)	.0063 (5)	.0124 (0)	0 (6,8)	.0373 (0)	0 (ALL)	0 (2,7,8,9)
Training Time (factor of NB time)	3.43	152.71	124.04	1	12.06	316.51
Overall Quality	**	*	***	*	***	***
Ease of Implementation	***	**	**	***	***	*
Ranking	4	6	2	5	1	3

Table 2: Results Classification (*** = good, ** = ok, * = poor)

References

1. Bezdek, James C. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media, 2013.
2. Blömer, Johannes, and Kathrin Bujna. "Simple Methods for Initializing the EM Algorithm for Gaussian Mixture Models." arXiv preprint arXiv:1312.5946, 2013.
3. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
4. Hsu, Chih-Wei, and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." *Neural Networks, IEEE Transactions on* 13.2 (2002): 415-425.
5. K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for l1-regularized logistic regression," J. Mach. Learning Res., vol. 8, pp. 1519–1555, 2007.
6. Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Cambridge: Cambridge university press, 2008.
7. Nongpiur, Monisha E., Benjamin A. Haaland, David S. Friedman, Shamira A. Perera, Mingguang He, Li-Lian Foo, Mani Baskaran, Lisandro M. Sakata, Tien Y. Wong, and Tin Aung. "Classification algorithms based on anterior segment optical coherence tomography measurements for detection of angle closure." Ophthalmology 120, no. 1 (2013): 48-54
8. Tjoa, Steven K., and KJ Ray Liu. "Multiplicative update rules for nonnegative matrix factorization with co-occurrence constraints." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.