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

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 plan 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 misclassification rates 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. We may also pursue sensitivity analysis to evaluate the consistency of performance in these areas.

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.

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 will be considered in this project. Following algorithms shall be compared:

- k-means clustering algorithm: Starting with a given number of random cluster centroids each data point is assigned to the closest centroid in terms of the ℓ_2^2 norm. The centroids are then updated as the mean of the data points assigned to it and the data points are assigned to the new closest centroid. This is repeated until convergence. The result is a hard clustering.
- k-median clustering algorithm: Starting with a given number of random cluster centroids 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 (less sensitive with respect to outliers than the mean) and the data points are assigned to the new closest centroid. This is repeated until convergence. The result is a hard clustering.
- Fuzzy c-means clustering algorithm: Fuzzy c-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. The centroid is then updated as the mean of all points weighted by their degree of belonging to this centroid. The result is a soft clustering.
- EM clustering algorithm for a principle components Gaussian mixture model: This algorithm combines the basic EM Algorithm for a Gaussian Mixture Model with a PCA 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 one further applies a spectral decomposition 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.
- Non-negative matrix factorization clustering algorithm: Using non-negative matrix factorization one approximates the data matrix X (n x m) as the product of a matrix A (n x k) and a matrix S (k x 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. The result is a soft clustering. (Tjoa and Liu 2010)

The algorithms are compared with respect to running time, clustering quality (visual inspection of cluster means and evaluation of misclassification rates based on the labels) as well as ease of implementation.

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.

Expected Results

We are testing many different methods and so we will limit our expectations to broad statements. We expect that SVM and Lasso multinomial regression will be the best classification methods. Naïve Bayes method will not perform well because the assumption of conditional independence is not likely to hold. The neural network method is difficult to implement but, if it is tuned properly, may perform exceedingly well.

In clustering, we expect the EM algorithm and the non-negative matrix factorization to perform the best. The k-means, k-median, and fuzzy c-means are likely not to perform as well due to the large dimensions of the problem. The EM algorithm also has a tuning parameter in the PCA covariance aspect which gives it a slight advantage in performance.

References

1. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
2. 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.
3. 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.
4. 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
5. Tjoa, S. K., & Liu, K. R. (2010, March). Multiplicative update rules for nonnegative matrix factorization with co-occurrence constraints. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 449-452). IEEE.