

Maximum Entropy Density Estimation Using a Genetic Algorithm

Parag C. Pendharkar, Penn State Harrisburg (pxp19@psu.edu)
James A. Rodger, Indiana University of Pennsylvania (jrodger@iup.edu)

Abstract

Several unsupervised learning algorithms, neural networks, and support vector machine based classification and clustering approaches are kernel-based, and require sophisticated algorithms for density estimation. The density estimation problem is a nontrivial optimization problem and most of the existing density estimation algorithms provide locally optimal solutions. In this paper we use an entropy maximizing approach that uses global search genetic algorithm to estimate densities for a given data set. Unlike the traditional local search approaches, our approach uses global search and is more likely to provide solutions that are close to global optimum. Using a simulated dataset, we compare the results of our approach with the maximum likelihood approach.

1. Introduction

Density estimation is a popular problem in data mining literature, and there are several data mining problems that lend themselves to density estimation (Tan et al., 2006). For example, several cluster analysis problems can be considered as multiple-density estimation problems where each probability density function represents a cluster. A few classification approaches such as radial basis function neural networks and support vector machines require transformation of non-linearly separable inputs vectors from high dimension to low dimensional vectors in feature space so that the inputs are linearly separable. Density or kernel estimation plays a vital role in this transformation of high dimensional input vectors into the low dimension feature space vectors (Haykin, 1999).

A density estimation problem can be defined as follows. Given a data set of size M of n -dimensional vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ that are assumed to be obtained from an unknown probability distribution, estimate the probability density function (pdf) $p(\mathbf{x}_j)$, where $j = \{1, 2, \dots, M\}$, that maps a vector \mathbf{x}_j to the interval $[0, 1]$. A density estimation problem is solved using parametric, non-parametric and semi-parametric approaches. Parametric approaches assume a particular data distribution and fit a pdf on the given dataset. Non-parametric approaches use data binning and histogram fitting, whereas the semi-parametric approaches use maximum likelihood and expectation maximization approaches (Miller and Horn, 1998). Parametric approaches are relatively simple and don't fare well particularly when the assumptions they make are violated. Non-parametric approaches face problems of computational complexity (Bishop, 1995). Semi-parametric approaches combine the advantages of both parametric and non-parametric approaches, and are more popular (Bradley et al., 1999).

Semi-parametric approaches are computationally efficient and provide locally optimal solutions. A popular semi-parametric density estimation approach is expectation maximization (EM) algorithm. The EM algorithm uses Gaussian mixture model and

maximum likelihood approach to estimate pdf for a given data set. The EM algorithm is well known to converge efficiently to local optimum (Fayyad et al., 1998).

Researchers from statistics and economics communities have recently shown interest in using maximum entropy (ME) measures for density estimation problems. For example, Miller and Horn (1998) illustrate an encoding and decoding approach based stochastic gradient ascent method for entropy maximization to estimate probability densities. Wu (2003) propose a sequential updating procedure to compute maximum entropy densities subject to known moment constraints. Wu and Stengos (2005) use a ME density approach to estimate error distribution in regression models. All researchers have reported good results with the use of ME measures. Wu and Stengos (2005) argue that ME densities have simple function forms that nest most of the commonly used pdfs including the normal distribution, which is considered as a special case as opposed to a limiting case. All the procedures proposed and used in Wu (2003) and Wu and Stengos (2005) studies converge to local optimum and suffer from similar computational issues as that of EM algorithm.

While the ME approach for density estimation has shown promise, current approaches cannot be directly applied to clustering problems in data mining. The approaches suggested by Miller and Horn (1998), Wu (2003) and Wu and Stengos (2005) either assume knowledge of data distribution or prior knowledge of movement constraints. Since a typical application of density estimation in data mining literature is likely to an unsupervised learning problem, such prior knowledge may not be available.

In this paper, we use a global search heuristic genetic algorithm for density estimation. Genetic algorithms (GAs) are population based parallel search techniques that are likely to find “heuristic” optimal solutions to optimization problems. Unlike gradient search algorithms used in EM and ME density estimation, which are likely to get stuck in local optimum, GAs are likely to provide solutions that are close to global optimum. We use GAs to estimate pdfs on a simulated dataset using both the maximum likelihood (ML) formulation and the ME formulation.

The rest of the paper is organized as follows. In section 2, we describe a semi-parametric density estimation problem and describe ML and ME objectives. In section 3, we describe our simulated data and experiments. In section 4, we conclude our paper with a summary.

2. Semi-Parametric Density Estimation

A typical semi-parametric density estimation consists of a finite mixture of density model of k probability distributions, where each of k distributions represents a cluster (Han and Kamber, 2006). The value of k is usually less than the number of data points. We use a Gaussian mixture model where true pdf is considered to be a linear combination of k *basis functions*. The basis functions are conditional probability densities $p(\mathbf{x}|i)$, which are linearly combined using the prior probabilities, $P(i)$, to represent the true pdf $p(\mathbf{x})$ as follows.

$$p(\mathbf{x}) = \sum_{i=1}^k p(\mathbf{x} | i)P(i).$$

For a Gaussian mixture model, the k components have Gaussian distributions with a mean vector for i^{th} component $\boldsymbol{\mu}^i \in \mathfrak{R}^n$ and a $n \times n$ covariance matrix of $\Sigma^i = (\sigma^i)^2 \times I$. The symbol I is the identity matrix. Thus, the components of basis functions are given by:

$$p(\mathbf{x} | i) = \frac{1}{(2\pi(\sigma^i)^2)^{\frac{n}{2}}} \times e^{\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}^i\|^2}{2(\sigma^i)^2}\right)}.$$

Using Bayes theorem and ML approach, Gaussian mixture model translates into following non trivial optimization problem with multiple local minima (see Bradley et al., 1999).

$$\underset{P(i), \boldsymbol{\mu}^i, \sigma^i}{\text{Minimize}} \quad - \sum_{s=1}^M \left(\log \left(\sum_{i=1}^k \frac{1}{(2\pi(\sigma^i)^2)^{\frac{n}{2}}} \times e^{\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}^i\|^2}{2(\sigma^i)^2}\right)} \times P(i) \right) \right).$$

When entropy based density estimation approach is used then the priors, mean vectors and covariances that maximize the following expression will provide the best estimate of the true pdf $p(\mathbf{x})$ (Cheng et al., 1999).

$$\underset{P(i), \boldsymbol{\mu}^i, \sigma^i}{\text{Maximize}} \quad - \sum_{s=1}^M \left(\left(\sum_{i=1}^k \frac{1}{(2\pi(\sigma^i)^2)^{\frac{n}{2}}} \times e^{\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}^i\|^2}{2(\sigma^i)^2}\right)} \times P(i) \right) \log \left(\sum_{i=1}^k \frac{1}{(2\pi(\sigma^i)^2)^{\frac{n}{2}}} \times e^{\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}^i\|^2}{2(\sigma^i)^2}\right)} \times P(i) \right) \right).$$

Like the ML approach, ME approach is also a nontrivial optimization problem with multiple local maxima.

We use genetic algorithm for optimizing the non-trivial ML and ME optimization problems. GAs are general purpose stochastic parallel search approaches that are used for optimization problems (Goldberg, 1989). We use floating point representation for our research. All the optimization variables are represented as genes in a population member. We use single point crossover and single gene mutation.

3. Simulated Data and Experiments

We generate a two attribute simulated dataset using simulations of three different normal distributions. We generate our data using three normal distributions with means of -1, 0, and 1. The standard deviations for all the distributions are considered equal to one. These three distributions represent three clusters. Figures 1 and 2 illustrate the data distributions and their contour plots. We generated 20 data points for each distribution with a total of 60 data points.

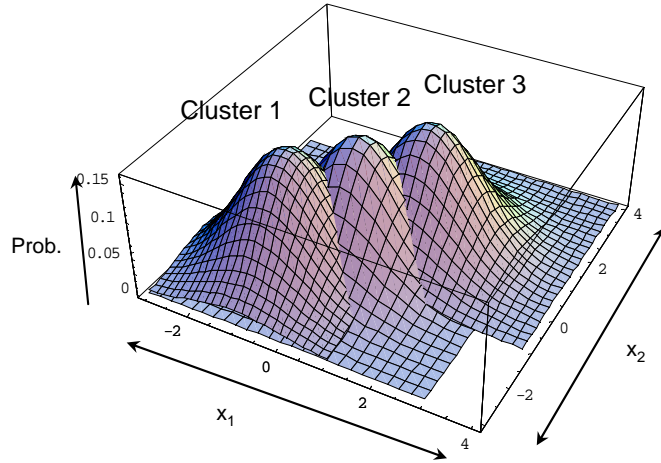


Figure 1: The data distributions for simulated dataset with three clusters

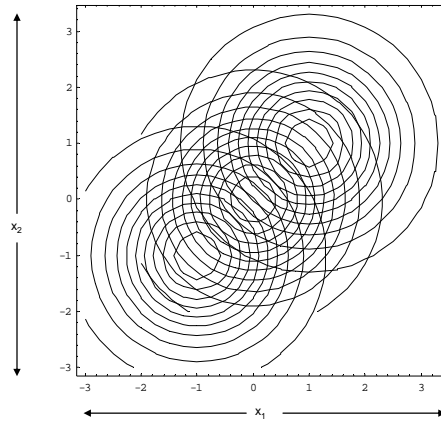


Figure 2: A contour plot illustrating the expected overlapping between clusters

We test the accuracy of our ML and ME optimization on our simulated dataset. Since the problem is that of cluster analysis, our data input vector contained the tuple $\langle x_1, x_2 \rangle$ and we specified the value of $k=3$ in our experiments. We use the GA procedure to optimize the ML and ME functions. The GA parameters were set after initial experimentation as follows. Mutation rate was set to 0.1, crossover rate was set to 0.3 and terminating iterations condition was set to 1000 learning generations. The cluster assignment for each data point in the sample was conducted after the optimization of the function and finding the maximum value of the likelihood determined by the following formula.

$$\arg \max_i \left(\frac{1}{\left(2\pi(\sigma^i)^2\right)^{\frac{n}{2}}} \times e^{-\left(\frac{\|x-\mu^i\|^2}{2(\sigma^i)^2}\right)} \times P(i) \right).$$

Table 1 shows the cluster means and standard deviations obtained for ME and ML objective functions. The cluster assignments from the algorithms were compared with the actual distribution from which a data point was generated to compute correct classification.

Table 1: Cluster Means and Standard Deviations for the two Objective Functions

Technique	Cluster 1			Cluster 2			Cluster 3			Percent Correct
	μ_1	μ_2	σ	μ_1	μ_2	σ	μ_1	μ_2	σ	
ME	1.00	0.54	0.60	0.24	0.83	0.48	0.93	1.39	1.44	53.33%
ML	0.76	0.52	0.29	0.79	0.54	0.32	0.78	0.59	1.44	45%

The preliminary results of our experiments indicate that ME approach, when compared to ML, appears to fare well. The standard deviations of ML approach are lower than or equal to ME approach and it appears that ML approach is overly conservative.

4. Conclusions

We used a heuristic and global search genetic algorithm for a non-trivial density estimation problem. The objective function for the density estimation problem was derived from the ME and ML approaches. Using a simulated data set generated with two attributes and three different normal distributions, we tested our genetic algorithm approach. Our results indicate that ME is a better objective function for density estimation problem.

References

- Bishop, C.M. *Neural Networks for Pattern Recognition*, Oxford University Press, New York, NY, 1995.
- Bradley, P.S., Fayyad, U. M., Mangasarian, O. L. Mathematical programming for data mining: Formulations and challenges, *INFORMS Journal on Computing*, 11(3), pp. 217-238, 1999.
- Cheng, C-H, Fu, A. W-C, Zhang, Y. "Entropy-based Subspace Clustering for Mining Numerical Data, *Proceedings of 5th International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pp. 84-93, San Diego, CA, 1999.

Fayyad, U. M., Reina, C. A., Bradley, P. S. Initialization of iterative refinement clustering algorithms, *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD98)*, pp. 194-198. AAAI Press, Menlo Park, CA, 1998.

Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

Han, J. and Kamber, M. *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2006.

Haykin, S. *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, New Jersey, 1999.

Miller, G., and Horn, D. Probability density estimation using entropy maximization, *Neural Computation*, 10, pp. 1925-1938, 1998.

Tan, P-N, Steinbach, M., Kumar, V. *Introduction to Data Mining*, Pearson Addison Wesley, Boston, MA, 2006.

Wu, X., and Stengos, T. Partially adaptive estimation via the maximum entropy densities, *Econometrics Journal*, 8, pp. 352-366, 2005.

Wu, X. Calculation of maximum entropy densities with application to income distribution, *Journal of Econometrics*, 115, pp. 347-354, 2003.