Classification Society 2008 Meeting
Sample of Abstracts

Multivariate Density Estimates using Geometric Methods

Leonard B Hearne
University of Missouri at Columbia
hearnel@missouri.edu

In this talk I focus on two geometric density estimators. These estimators use the Delaunay tessellation or its geometric dual, the Voronoi diagram, to partition the support for an estimator into tiles. The proportion of probabilistic mass from observations on each tile divided by the content of the tile is used to estimate the probability density on each tile. Under suitable regularity conditions geometric density estimators can be shown to be consistent multivariate estimators. For both types of tessellations, the level of density specificity is proportional to the number of tiles in the tessellation.

Voronoi tiles are generated by an observation and are the set of points in the support space closer to an observed value than to any other observation. The support for an estimator using the Voronoi diagram is the whole space and if an observation is on the convex hull of the observations then the content of the generated tile is infinite so the density estimate on this tile is zero.

The Delaunay tessellation has observations as vertices of simplexes such that the circumscribed spherical cover of each tile does not contain any points in the set of observation in its interior. The support for the estimator is the content of the convex hull of the set of observations. The probabilistic mass of vertices is assigned to adjacent tiles in the tessellation and the support for the estimator is the contents of the convex hull of the observations.

To refine a density estimate re-sampling methods can be employed. This partitions the support for the estimator into successively smaller tiles. With both the Voronoi and Delaunay tessellations the resulting refined density estimator is biased. By truncating the Voronoi tiles generated by points on the convex hull or by allowing some probability mass to be allocated beyond the convex hull in a Delaunay tessellation, these refined density estimators can be made unbiased and consistent.

This work has application in a broad class of multivariate estimation settings where geometric methods can be employed.

Single linkage cluster analysis as a tool for hypervariate outlier detection prior to classification

Peter L. Flom
BrainScope, Inc
peterflomconsulting@mindspring.com

A commonly cited drawback of single-linkage cluster analysis is its tendency to chain, that is, to have many observations that join clusters very late in the hierarchical process.

The detection of outliers is a complex problem, even when the data set has only one or a few variables. Many graphical procedures (e.g. Andrews curves, Chernoff faces) may be useful when the data are multivariate, provided that the number of variables is not very large. However, when there are hundreds, or thousands, of variables, these methods are not practical.

We propose using the chaining property of single-linkage cluster analysis to identify observations that may be outliers in a hypervariate space, even if they are not outliers in any univariate or bivariate space. We illustrate this method with a set of quantitative electroencephalographic (QEEG) data , and show how removal of outliers can improve classification.

Simultaneous Cancer Classification and Gene Selection with Bayesian Nearest Neighbor Method: An Integrated Approach

SOUNAK CHAKRABORTY
University of Missouri-Columbia
chakrabortys@missouri.edu

Since most of the cancer treatments comes with certain degree of toxicity it is very essential to identify a cancer type correctly and then administer the relevant therapy. With the arrival of powerful tools like gene expression microarrays the cancer classification basis is slowly changing from morphological properties to molecular signature. Thus we are hoping to reduce the classification error. The main challenge working with gene expression microarray is there are huge number of genes to work with. And out of them only a small fraction of are actually relevant for differentiating between types of cancer.

In this paper we built a Bayesian nearest neighbor model equipped with an integrated variable selection technique to overcome this challenge. Our developed classification and gene selection model is able to accurately classify different cancer types and simultaneously select relevant genes. Our proposed model is completely automatic in the sense that it adaptively picks up the neighborhood size and important genes. The method is successfully applied to four well known data sets. When compared with other “off the shelf” classification methods like random forest, support vector machine, neural network, and k nearest neighbor, for all the data sets studied our method produced highly competitive if not better results. Also adaptive gene selection technique helps us to monitor the selected genes by our model and study their biological relevance.

The Ultrametric Topology Perspective on Analysis of Massive, Very High Dimensional Data Stores

Fionn Murtagh

Science Foundation Ireland, Dublin, Ireland, and Department of Computer Science, Royal Holloway, University of London

An ultrametric topology formalizes the notion of hierarchical structure. An ultrametric embedding, referred to here as ultrametricity, is implied by a hierarchical embedding. Such hierarchical structure can be global in the data set, or local. By quantifying extent or degree of ultrametricity in a data set, we show that ultrametricity becomes pervasive as dimensionality and/or spatial sparsity increases. This leads us to assert that very high dimensional data are of simple structure. We exemplify this finding through a range of simulated data cases. We discuss also application to very high frequency time series segmentation and modeling. Other applications will be described, in particular in the area of textual data mining.

References

[1] F. Murtagh, On ultrametricity, data coding, and computation, Journal of Classification, 21, 167-184, 2004.

[2] F. Murtagh, G. Downs and P. Contreras, "Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding", SIAM Journal on Scientific Computing, 30, 707-730, 2008.

[3] F. Murtagh, The remarkable simplicity of very high dimensional data: application of model-based clustering, submitted, 2007.

[4] F. Murtagh, Symmetry in data mining and analysis: a unifying view based on hierarchy, submitted, 2007.

[5] F. Murtagh, “On ultrametric algorithmic information”, Computer Journal, in press, 2008. (Online access 9 Oct. 2007.)

Permanent Classification Model with Application to Microarray Analysis

Jie Yang
University of Illinois at Chicago
jyang06@math.uic.edu

This talk will introduce a new statistical model based on a permanent process for supervised and unsupervised classification problems. Unlike most research works in the literature, the new model assumes only exchangeability instead of independence on observations. Regardless of the number of classes or the dimension of the feature variables, the model may require only 2-3 parameters for fitting the covariance structure within clusters. It works well even if the class occupies non-convex, disjoint regions, or regions overlapped with other classes in the feature space. The application to DNA microarray analysis leads to some interesting discoveries.

Nonparametric Clustering on Mixtures of Longitudinal/Functional Data

Haiyan Wang
Kansas State University
hwang@ksu.edu

In this talk, I will present a method for effectively detecting unknown clusters in high dimensional longitudinal or functional data. Examples of such data include gene expression levels measured over time from microarray experiments, functional magnetic resonance imaging (fMRI), mass spectrometry data from proteinomics, lipidomics etc. We define clusters through the unknown high dimensional multivariate distributions of all observations. Kullback-Leibler information and Mahalanobis generalized squared distance can fail to provide meaningful measure of distance between distributions in high dimensional settings. We propose a new similarity measure and a clustering algorithm to effectively differentiate among high dimensional populations. The algorithm produces invariant results under monotone transformations of data and does not require users to specify the number of clusters. Simulations show that PCLUST significantly outperforms 9 other popular algorithms in both clustering accuracy and robustness. An application in identifying biomarkers using time course gene expression data from Arabidopsis in response to environmental stresses is illustrated.

The talk is based on my joint work with James Neill, Forrest Miller and George von Borries.

Elemental Set Methods

David Banks
Duke University
banks@stat.duke.edu

Elemental set methods are an old idea for finding simple structure hidden in complex noise. This talk describes research on extending the method to higher dimensions, and contrasts its performance with Bayesian mixture modeling. The method is applied to linear, nonlinear, and nonparametric regression, as well as multidimensional scaling.

Visualizing Hierarchical Cluster Structure via Linkage Algorithms

Rebecca Nugent
Carnegie Mellon University

The goal of clustering is to identify distinct groups in a data set and assign a group label to each observation. To cast clustering as a statistical problem, we regard the data as a sample from an unknown density p(x). To generate clusters, we estimate the properties of p(x) with either parametric (model-based) or nonparametric methods. In model-based clustering, we assume that groups in the population correspond to mixture components in the density estimate; in nonparametric clustering, they correspond to the density estimate's modes. In contrast, the algorithmic approach to clustering (linkage methods, spectral clustering) applies an algorithm, often based on a distance measure to data in m-dimensional space. Clusters are extracted heuristically. We propose to combine the strengths of the different clustering approaches to visualize the (possibly hierarchical) cluster structure.

After briefly introducing generalized single linkage, we sketch the theoretical connections between nonparametric clustering and linkage algorithms. We propose the utilization of a linkage algorithm with a minimum density similarity measure to visualize the hierarchical structure of the modes (or components) of the density estimate. The resulting dendrogram can then be used as a tool to subjectively prune or merge clusters. Finding the minimum density between observations is an optimization problem with no closed form solution. Our similarity measure can be approximated; however, we will present an algorithm employing Taylor series expansion bounds that allows us to generate a dendrogram without requiring the exact similarities. We then will discuss the advantages and disadvantages of single and complete linkage and a few possible minimum density similarity measures. A dendrogram pruning algorithm will be introduced; results will be presented for real and simulated data sets.

Augmenting Model-Based Clustering with Generalized Linkage Methods

Nema Dean
University of Glasgow

The fundamental assumption made by model-based clustering (Fraley and Raftery 1998) is that the groups or sub-populations underlying the data have (multivariate) Gaussian distributions, giving the overall population a finite mixture model distribution. A common assumption additionally made, is that the number and type of components found to best fit the data are a good estimate of the number and type of true groups in the data. Given the shape assumptions implicit in the choice of Gaussian distributions - elliptical, symmetric contours - in cases of skewed, curved or more generally complex-shaped groups, the equivalence of the mixture model components and the underlying groups may be false.

Since general continuous densities can be modelled arbitrarily well by mixtures of Gaussian densities, the mixture model chosen may still be a good estimate of the density of the data but it is likely that more than one component is identified with each group. Generalized linkage methods (Stuetzle and Nugent 2007) use density estimates to create density-based similarity (or dissimilarity) measures which can then be used as a replacement for Euclidean (or other types of) distance in hierarchical agglomerative methods. Utilizing the resulting model-based clustering density estimate we can apply the resulting dendrogram to visualize the hierarchical structure of the components of the mixture model and make decisions about combining components to estimate groups. Since it is difficult to easily summarize information about complex shaped groups, offering a summary that is essentially a subset of components of the original mixture model with means and covariance matrices is an attractive alternative.

Visualizing and Clustering Students' Skill Sets in a Unit Hyper-cube

Beth Ayers
Carnegie Mellon University

In educational research, a fundamental goal is identifying which skills students have mastered, which skills they have not, and which skills they may be in the process of mastering and then grouping students into common skill set profiles. Combining a student response matrix that indicates which questions the students attempted and whether or not they answered them correctly with an expert-elicited assignment matrix of the K possible skills required for the questions, students are mapped via a capability matrix to a skill space lying on the unit hyper-cube. Each skill corresponds to a dimension of the hyper-cube. Along each dimension, a one indicates a mastered skill, zero indicates a skill yet to be mastered, and 0.5 indicates a skill that is partially mastered or a skill about which we are uncertain. Then each corner of the unit hyper-cube is one of the natural 2^K possible skill set profiles.

In practice, students may only see a limited number of questions, questions may require combining several skills, and some skill set profiles may be completely absent from the data. We use the capability matrix to estimate the number and type of different groups of underlying skill sets present in the student population, to assign each student to a group, and to summarize the group structure. To accomplish these tasks, we combine both model-based clustering and a variation of k-means with nonparametric visualization techniques. Model-based clustering allows us to remove the assumption of a fixed number of skill sets and identifies small highly concentrated groups of students. Our variation of the k-means algorithm uses as natural starting centers the corners of the hyper-cube and allows for empty clusters (skill sets or corners with no students). For both methods, linkage algorithms using a minimum density distance are utilized to visualize the (possibly hierarchical) skill set structure. Results are compared across a series of simulated and real data sets.