Multivariate clustering and classification

Eric Feigelson

Statistical approach to unsupervised clustering

In **unsupervised clustering** of a multivariate *nxp* dataset, the number, location, size and morphology of the data groupings is unknown. There is no `prior knowledge' of classes.

Nonparametric clustering algorithms:

- Agglomerative hierarchical clustering
- K-means partitioning
- Density-based clustering

Parametric clustering algorithms:

- Normal mixture models

Nonparametric unsupervised clustering is a very uncertain enterprise, and different algorithms give different outcomes without mathematical guidance (e.g. there is no likelihood to maximize or stopping criterion to choose number of clusters). Results should be viewed with great caution for scientific inference.

Parametric unsupervised clustering lies on a stronger foundation (e.g. there is a likelihood to maximize, and BIC/AIC for model selection). But it assumes the clusters in fact follow the chosen parametric form.

Agglomerative hierarchical clustering

1. Construct the distance matrix $d(\mathbf{x}_i,\mathbf{x}_j)$ for the dataset, assuming a distance metric (e.g. Euclidean with standardized variables). Call each point a `cluster'.
2. Merge two clusters with the smallest `distance'. Several common choices for measuring the `distance' between a cluster and a new data point:

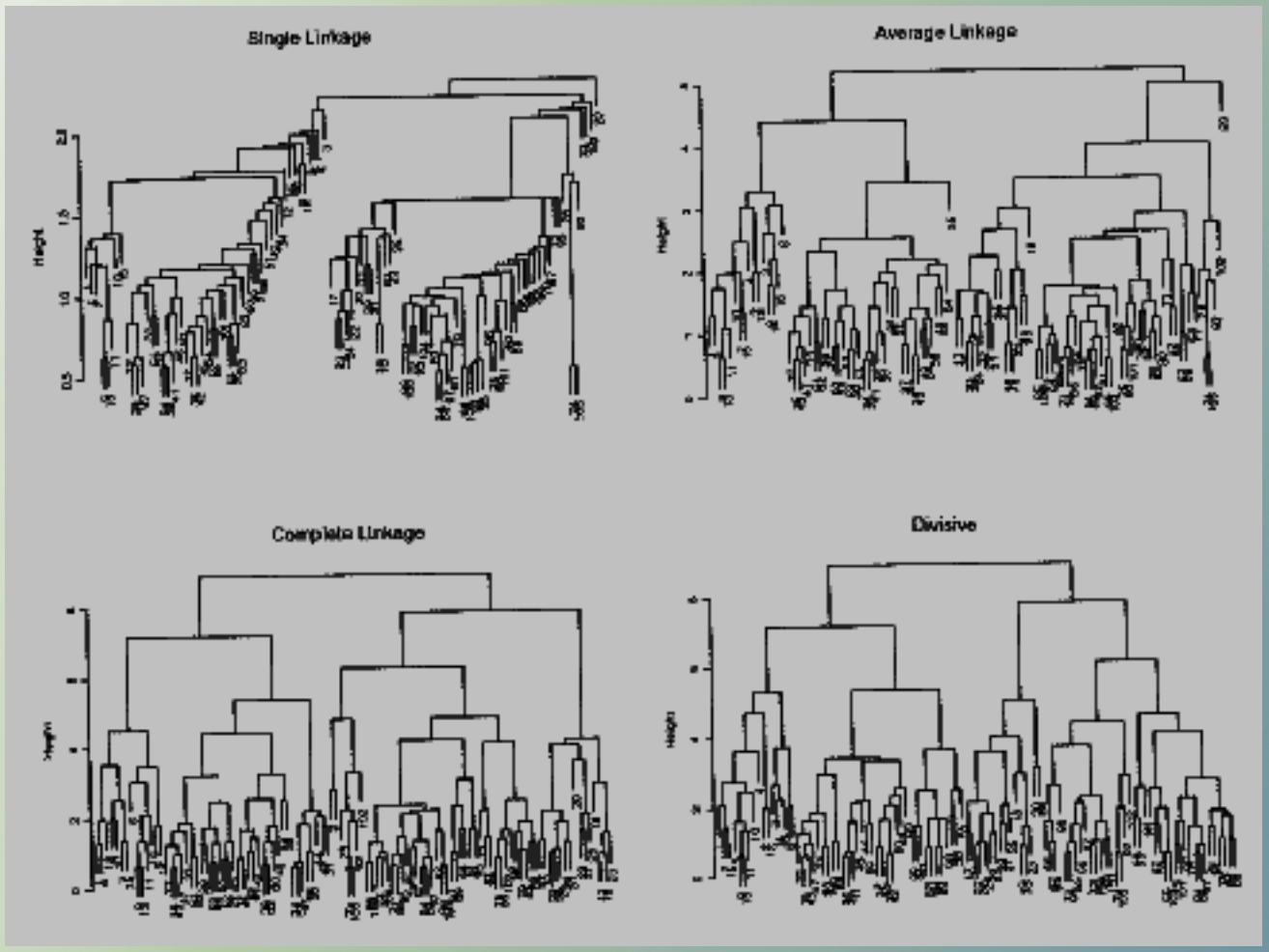
   - Minimum distance between any constituent point of the cluster and the new point = **single linkage clustering**. This procedure is equilvalent to `pruning' the **minimal spanning tree** of the multivariate dataset. This is the astronomers' friends-of-friends or percolation algorithm. This method is vulnerable to spurious `chaining' of distinct clusters into elongated superclusters, and is ***not recommended*** by statisticians.

   - Average distance between the constituent points of the cluster and the new point = **average linkage clustering.** This often gives an intermediate outcome but is scale-depenent.

   - Maximum distance between any constituent point of the cluster and the new point = **complete linkage clustering**. This is a conservative procedure that tends to give hyperspherical clusters.

   - Minimize the intra-cluster variances (**W** matrix) = **Ward's minimum variance clustering**

The result of an agglomerative (or divisive) clustering procedure is a dendrogram, or tree, showing the membership of each cluster at each stage of the clustering. ***There is no mathematical basis for choosing where to cut the tree, and thereby establishing the true number of clusters present.*** Qualitatively, objects combined at greater `heights' in the dendrogram are more dissimilar.



Comparison of hierarchical clustering methods

Primate scapular shapes
N=105, p=7

A. J. Izenman
Modern Multivariate Statistical Techniques
(2008)

Density-based clustering

Computer scientists and statisticians have developed methods for cases familiar to astronomers: where several clusters may exist in a background of unclustered objects. The methods assign objects either to clusters or to an unclustered background. Unlike k-means, these methods do not need a specified number of clusters.

Friedman & Fisher (1998) **bump hunting (PRIM)**: progressively shrink hyperrectangles to increase the enclosed density of points with a preset minimum population. Remove box from the dataset, and repeat. User interaction permitted. Related to CART.

DBSCAN (density-based spatial clusters of applications with noise) by Ester et al. (1996). User specifies minimum population and maximum extent (`reach') of a cluster. A cluster is expanded based on a single-linkage criterion.

Others include **BIRCH, DENCLUE, CHAMAELEON, OPTICS, …**

Normal mixture models

These are parametric regression models where the multivariate dataset is assumed to consist of *k* multivariate normal (MVN) clusters.

Each cluster has a hyperellipsoidal morphology extending over the entire space with mean vector μ_j and covariance matrix Σ_j where j=1, …, p.

The model has *2kp+k+1* parameters: *k* means and *k* variances in *p* dimensions, *k* mixture weights, and *k* itself.

Parameters are estimated by MLE using the EM Algorithm:

- Seed values of k, μ_j , and lower bound to Σ_j must be provided
- E step: Calculate likelihood of each object lying in each cluster
- M step: Cluster μ_j and Σ_j are updated with weighted objects
- Iterate EM until likelihood (or MAP for Bayesian) is maximized
- Run for different *k* with model selection from BIC or bootstrap

Other techniques include use of **W** and **B** matrices, robust procedures, and penalties for roughness

Codes include EMMIX, MCLUST, and AutoClass
(http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass, Cheesman & Stutz 1995)

Concepts of classification

Here we consider situations where the multivariate dataset under study represents a new ***test set*** that is a mixture of classes that have been defined in advance. Either the properties of these classes are known exactly from some prior knowledge (e.g. astrophysical theory) or, more often, is estimated from ***training sets*** where the objects have measured properties and known classes. Classification is thus a ***supervised*** process based on prior experience.

The existence of prior knowledge of the number and location of the classes in *p*-space gives a huge advantage over unsupervised clustering.

As with clustering, some classification methods are parametric (assuming multivariate normal distributions within each class) while others are nonparametric without any assumption of class morphology in *p*-space.

Automated classification techniques are particularly important in ***wide-field astronomical surveys*** which, like a dragnet that collects all sorts of sea animals from the ocean, collect a wide variety of astronomical objects: stars, galaxies, active galactic nuclei (accreting black holes), etc.. These broad classes then need to be classified further if the datasets are sufficiently detailed (e.g. multiband photometry, spectroscopy).

1. Stars can be classified into OBAFGKMLTY spectral types, WD/MS/RG/SG luminosity classes, and for multi-epoch data tracing variability, into ~80 classes of variable stars

2. Galaxies can be classified into Hubble morphological types (E/S0/S/Irr), clustering environments, and by star formation activity

3. Active galactic nuclei can be classified into Seyfert 1-1.5-2, FR Class I/I radio galaxies, BL Lacs/blazars, LINERS

Recent and planned wide-field surveys include:

- Optical photometry: CRTS, PTF, ASAS, Pan-STARRS, VISTA, DES, LSST, …
- Optical spectroscopy: LAMOST
- X-ray: RASS, eROSITA
- Infrared: IRAS, MSX, Akari, WISE
- Radio: NVSS, FIRST, PKS, LOFAR, MWO

Enormous resources worldwide are being devoted to Big Data surveys that will require classification

Statistical Classification & Data Mining

A few mid/late-20th century statistical methods for classification were developed by statisticians, some assuming MVN distributions and others nonparametric.

But most methodological development has been nonparametric methods led by scholars in the computer science & engineering communities. These go under the rubrics of ***Data Mining*** and, for a large class of iterative methods, ***Machine Learning***. Data Mining often includes suites of methods that perform data visualization, characterization, regression, and classification.

Data mining resources for astronomers are rapidly emerging:

- *Data Mining and Machine Learning for Astronomy*, N. Ball & R. Brunner, 2010 *Intl J Mod Phys D* (http://arxiv.org/abs/0906.2173)
- *A User's Guide for Data Mining in Astronomy*, 2011, N. Ball & S. McConnell (http://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaKDDguide)
- **Advances in Machine Learning and Data Mining in Astronomy**, M. Way, J. Scargle, K. Ali & A. Srivastava (eds.), CRC Press 2012
- **Statistics, Data Mining and Machine Learning in Astronomy**, Z. Ivezic, A. Connolly, J. VanderPlas & A. Gray, 2013, forthcoming text with Python code (http://astroml.github.com/)

Parametric classifiers

Assigning new members to two preexisting MVN clusters (Wald, 1940s)

The dataset $\mathbf{X} \sim N_p(\mu, \Sigma)$ consists of two clusters with $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1, \text{and } \mathbf{S}_2$

A new object with location x_0 is assigned to Cluster 1 if

$$\mathbf{x}_0' \left(\frac{1}{\mathbf{S}_1} - \frac{1}{\mathbf{S}_2} \right) \mathbf{x}_0 + \left(\frac{\bar{\mathbf{x}}_1'}{\mathbf{S}_1} - \frac{\bar{\mathbf{x}}_2'}{\mathbf{S}_2} \right) \mathbf{x}_0 - \frac{1}{2} \ln \frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}$$
$$-\frac{1}{2}(\bar{\mathbf{x}}_1' \mathbf{S}_1^{-1} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2' \mathbf{S}_2^{-1} \bar{\mathbf{x}}_2) \geq -2 \ln \left(\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right)$$

where $c(1|2)$ is the `cost' of misclassifying an object into cluster 1 when it truly belongs in cluster 2, and p_1 is the prior knowledge of the fraction of objects lying in cluster 1. These play roles similar to Type 1 & 2 errors in hypothesis testing.

Linear discriminant analysis (Fisher 1930s)

LDA finds a linear combination of variables (a *p*-dimensional hyperplane) that maximally separates two classes with known MVN distributions. The separation is measured by the ratio of the between-cluster variance **B** and the within-cluster variance **W**. The maximum separation occurs for

$$Sep = \frac{\mathbf{B}}{\mathbf{W}} = \frac{\mathbf{a}^2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^2}{\mathbf{a}'(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{a}} \quad \text{where}$$

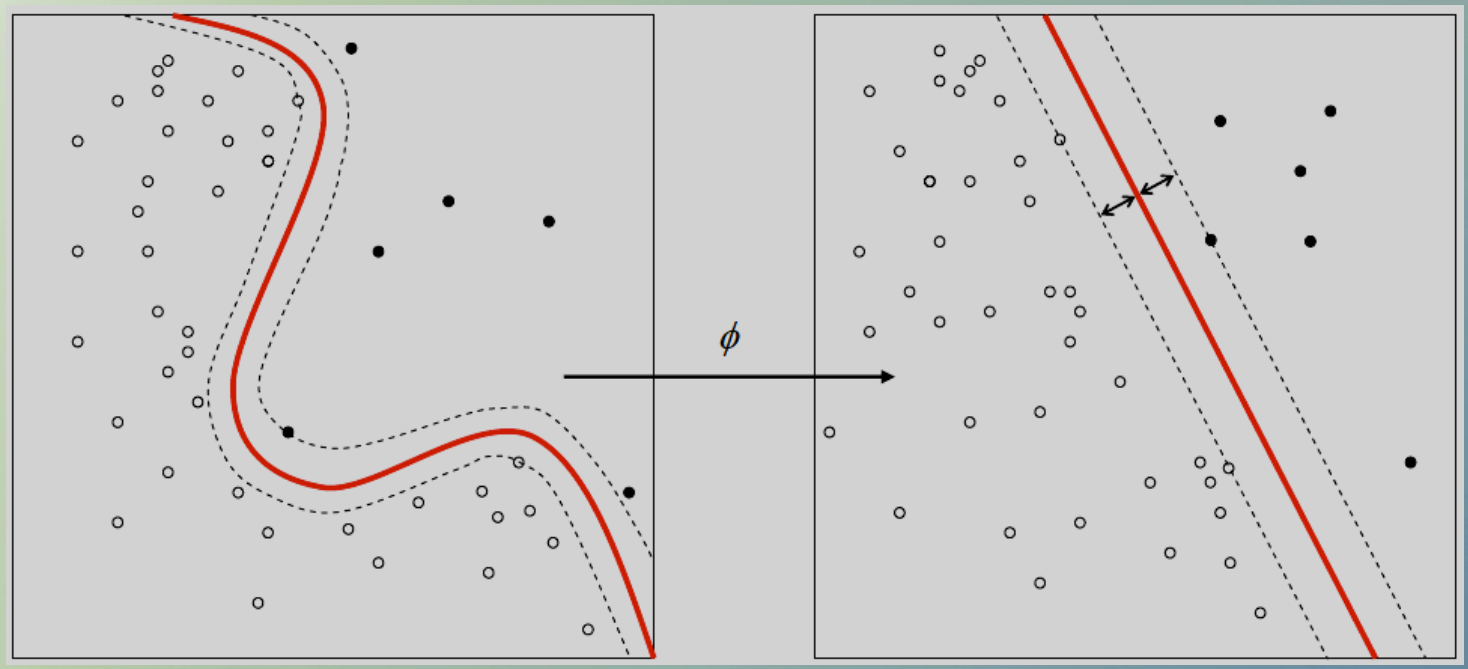
$$\mathbf{a} = \frac{\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1}{\mathbf{S}_1 + \mathbf{S}_2}.$$

where the vector a is perpendicular to the separating plane. The resulting separating plane can be used to understand the nature of the clustering, or can be applied for classification of new objects with unknown class.

LDA is often used today, and provides the foundation for numerous approaches to classification:

- Multiple discriminant analysis treats >2 classes
- Quadratic and polynomial discriminant analysis treat nonlinear problems
- Canonical discriminant analysis give reduction of dimensionality that emphasizes classification structure
- Its assumption of independence between the variables (use of covariance matrix with zero cross-terms) is equivalent to the **naïve Bayes classifier** which relaxes the assumption of MVN distributions.
- The **perceptron algorithm** related to LDA gives classifications for binary variable

Support Vector Machines (SVMs), developed by Vladamir Vapnik from the 1960-1990s, have emerged as extremely powerful generalizations of LDA and the perceptron. To treat cases where the hypercurve separating classes is nonlinear in *p*-space, the dataset is mapped by nonlinear functions onto a higher dimensional space where the classes can be separated by linear hyperplanes. The **support vectors** straddle the optimal hyperplane. Kernel density estimation (with polynomial or Gaussian kernels) plays an important role in the calculation that involves quadratic programming with Lagrangian multipliers. `Soft margins' allow the separation to have misclassifications.



http://www.youtube.com/watch?v=3liCbRZPrZA
http://www.jstatsoft.org/v15/i09/paper

Classification trees

Recall how unsupervised hierarchical clustering techniques construct a *dendrogram* from a multivariate dataset, where objects and subclusters that are `close' to each other (according to some distance metric and agglomeration algorithm) form branches of a tree where the `trunk' represents the full dataset and the `leaves' represent individual objects.

Recall also how astronomers often design heuristic decision rules for classification based on criteria like `color index > 0.4 mag' or `burst duration < 2 seconds'.

In 1963, Morgan & Sonquist proposed a **recursive** partitioning algorithm to construct decision trees for supervised classification. These were extensively developed from the 1970s-2000s by Leo Breiman at UC Berkeley. His methods are known as **Classification and Regression Trees (CART)**. Modern versions of CART often use the **ID3 or C4.5 algorithm** with tree reliability evaluated using the bootstrap-based **Random Forests** procedure.

CART

CART supervised classification procedure that constructs dendrograms for the training set where decisions are based on sequences of single-variable decision rules and the branching is designed to concentrate objects of a single class.

CART has important advantages:

- it does not depend on a distance metric (e.g. Euclidean distances)
- calculations are local with low memory requirements
- it is nonparametric (e.g. class shapes need not be MVN)
- it works for any combination of real, integer, categorical, or binary variables
- the same rules are used for small and large branches (i.e. recursive procedure)
- each data point falls into a unique terminal branch (node), and each terminal node has a unique set of rules (i.e. no branch crossings)
- it has objective mathematical procedures for constructing the full tree (leaves to trunk), pruning the tree, and evaluating the reliability of branches

However, CART does not give probabilities of membership, and it requires some user choices of technique and thresholds

Classification tree: outcomes assign objects to classes
Regression tree: outcome is a real number for a response variable

CART decision rules (choice of variable, value of split) minimizes the `impurity' of branching, with several measures of impurity in common use:

$$i(m) = \begin{cases} 1 - \max_j P_j & \text{misclassification impurity} \\ P_j P_k & \text{variance impurity} \\ -\sum_j P_j \log_2 P_j & \text{entropy impurity} \\ \frac{1}{2}\left[1 - \sum_j P_j^2\right] & \text{Gini impurity} \end{cases}$$

where P_j is the fraction of training set objects in the j-th class

Splitting stops, or a full tree is pruned, to some threshold level of impurity improvement or some penalty for model complexity.

Branch reliability can be evaluated by `votes' of trees constructed from many bootstraps of the training set, ***bagging***. An important variant of bagging is Breiman's ***Random Forest.*** Weak classifications can be weighted and combined, ***boosting.***

k-Nearest Neighbor classifiers

This is an extremely simple classification algorithm:

- Define a training set, test set, distance metric, and integer parameter *k*
- For each member of the test set, locate the *k* nearest neighbors of the training set. *k* plays a role similar to the bandwidth of kernel density estimation (KDE) or the window in local regression (e.g. LOESS).
- These points `vote', and the test set point class is set to be the most common class of the *k* neighbors. For two classes, majority wins.

As with bandwidth selection in KDE, *k* can be chosen to optimize some quantity. For classification, one may choose the *expected cost of misclassification,*

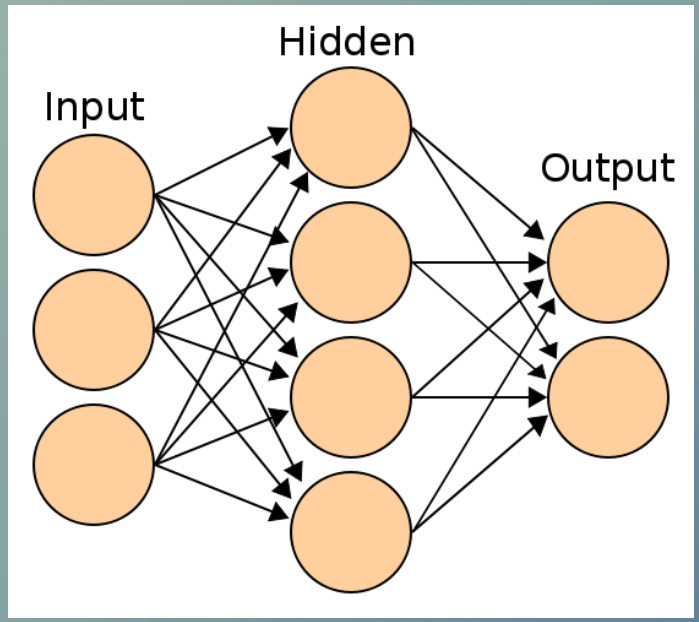
$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

k-nn classifiers are often used in machine learning; e.g. for optical character recognition. *k*-nn can be computationally expensive for Big Data, as accuracy generally increases with *k*.

Automated Neural Networks

ANNs are algorithms to find heuristic nonlinear rules for distinguishing classes in multivariate training datasets which can then be applied to test datasets. This is the most widely used data mining method in astronomy with ~700 papers since c.1990 accelerating to ~70/yr in 2012.

A 3- or 4-layered structure is created where the *nxp* data are inputs and the *p* classes are outputs. The intermediate `hidden' layers are weightings that probabilistically assign inputs to outputs (a generalization of the perceptron).



Hidden layer weightings are iteratively reset to improve classification using **back propagation**, a gradient descent procedure.

Many choices in network architecture, `activation functions' at the hidden nodes, optimality criteria (e.g. reducing the mean square error in classification), and stopping rules. Bayesian variants.

Convergence is not guaranteed.

Usually not possible to interpret the weightings … the proverbial `black box'.

Often highly effective for complex classification problems with large training sets.

Not advised for simple problems.

Final remarks

The word `classification' appeared in ~1/3 (2800/9000) of all astronomy papers published during 2012. Astronomers encounter endless problems where patterns are sought in heterogeneous data by placing objects into distinct classes.

Most astronomers still use heuristic procedures for classification. But a vanguard recognize the tremendous growth in quantitative methods for classification of multivariate datasets since the 1990s:

- If no prior knowledge on classes is available, then uncertain nonparametric, or focused parametric (e.g. MVN, isothermal ellipsoid) clustering methods are needed.

- If prior knowledge is available, then a vast suite of powerful supervised classification methods are available: SVMs, CARTs, boosting & bagging, k-NNs, ANNs