Data Mining: Multivariate clustering & classification with R

Eric Feigelson

NARIT-EACOA Summer Workshop on Astrostatistics & Astroinformatics
August 2019

Astronomical context

Astronomers have constructed classifications of celestial objects for centuries:

- Asteros (fixed stars) vs. planetos (roving stars) [Greece, >2Kyr ago]
- Luminosae, Nebulosae, Occultae [Hodierna, mid-17th c]
- Comet orbits & morphology [Alexander 1850, Lardner 1853, Barnard 1891]
- Stellar spectra: 6 classes (Secchi 1860s), 7 classes (Pickering/Cannon 1900-20s), 10 classes w/ brown dwarfs (Kirkpatrick 2005)
- Galaxy morphology: 6+3 classes (Hubble 1926)
- Supernovae: Ia, Ib, Ic, Iib, IIP, Iin (Turatto 2003)
- Active galactic nuclei: Seyfert gal, radio gal, LINERs, quasars, QSOs, BL Lac, blazars
- Protostars/PMS stars: Class 0, 0/I, I, II, III (Lada 1992, Andre 1993)

In nearly every case, these classes were created by well-argued, but subjective assessment of source properties.

In statistical parlance, the problem is called
unsupervised clustering of heterogeneous multivariate data

(Poor) clustering methods in astronomy

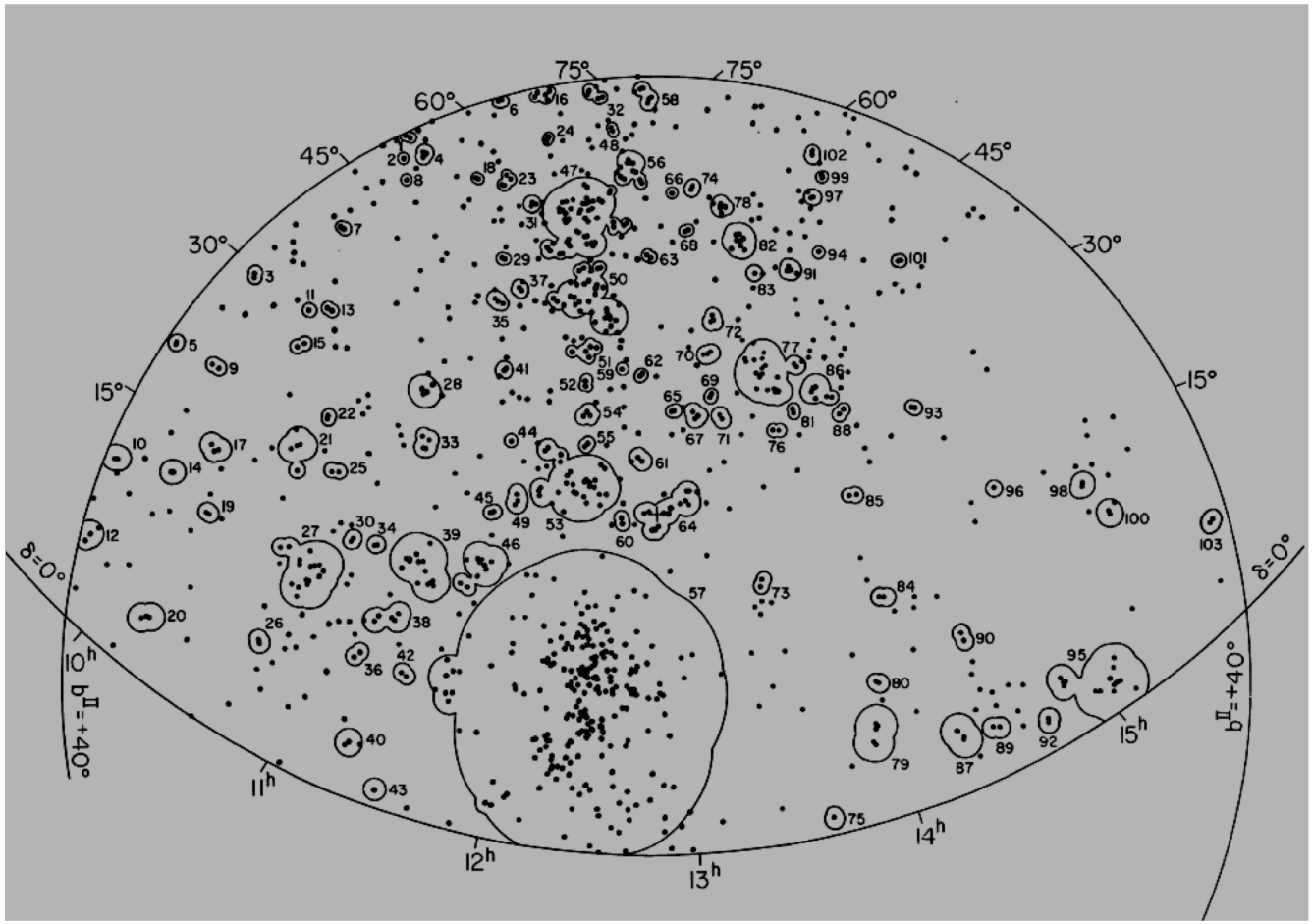
Deterministic decision trees

- Abell cluster richness class (Abell 1958)
- Young stellar objects with infrared colors $0.4<[5.8]-[8.0]<1.1$ and $0.0<[3.6]-[4.5]<0.8$ are classified as *Class II* (Allen 2004)

Percolation or `friends-of-friends' algorithm

1. Plot data points in a 2-dimensional diagram
2. Find the closest pair, and call the merged object a `cluster'
3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated

2. Percolation or `friends-of-friends' algorithm



Turner & Gott
Groups of Galaxies I
A Catalog ApJS 1976

In statistics, this is `single linkage hierarchical clustering'

Statistical approach to clustering

In **unsupervised clustering** of a multivariate *n x p* 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 ~ Friends-of-friends
- K-means partitioning
- Density-based clustering

Parametric clustering algorithms:

- Mixture models

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:

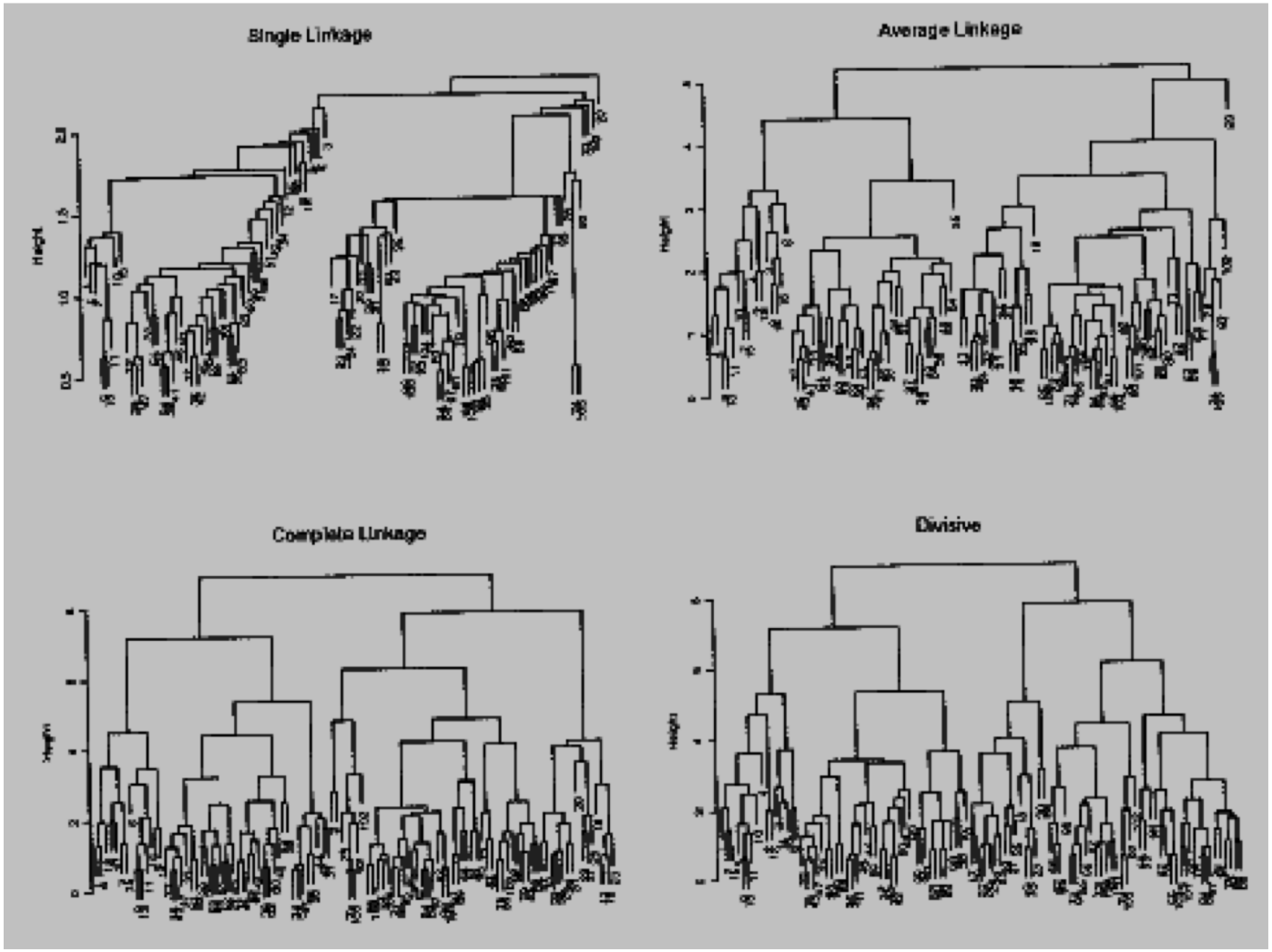
   - Minimum distance between any constituent point of the cluster and the new point = **single linkage clustering**. This procedure is equivalent to `pruning' the **minimal spanning tree** of the multivariate dataset. This is the astronomers' friends-of-friends or percolation algorithm. This method is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

   - 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-dependent.

   - 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 (tr **W**) = **Ward's (1963) 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)

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

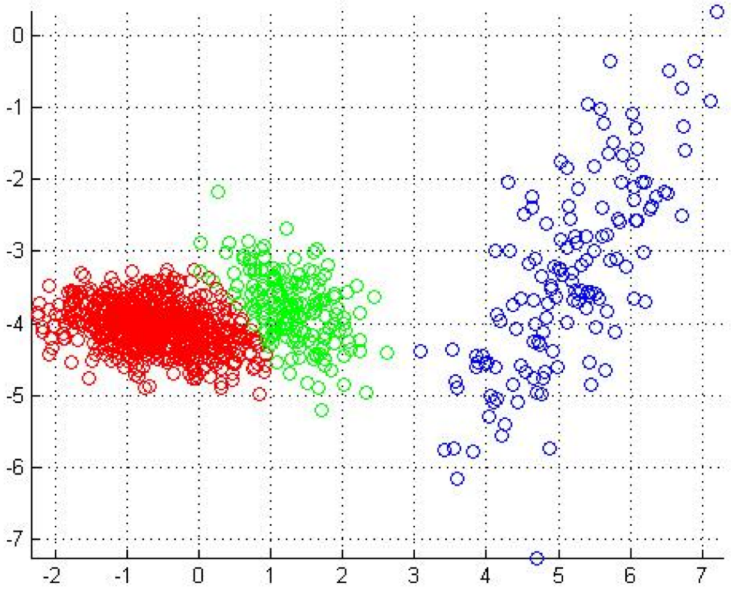
Parametric unsupervised clustering lies on a stronger foundation: there is a likelihood to maximize, and tools like BIC/AIC for model selection. But it assumes the cluster shapes in fact follow the chosen parametric form.

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.



Concepts of [supervised] classification

The multivariate dataset under study represents a new ***test set*** that is a mixture of classes that have been defined in advance, either from astrophysical theory or ***training sets***. The prior knowledge of the number, location & morphology of the classes in *p*-space gives a huge advantage over unsupervised clustering.

As with clustering, some classification methods are parametric assuming multivariate normal (MVN) distributions within each class (mixure models), while others are nonparametric. Methods often labeled ***data mining*** or ***machine learning***.

Automated classification techniques are particularly important in ***wide-field astronomical surveys*** which collect a wide variety of astronomical objects: stars, galaxies, active galactic nuclei.

Wide-field surveys include: Optical CRTS, PTF, ASAS, Pan-STARRS, VISTA, DES, LSST, LAMOST; **X-ray** RASS, eROSITA; **Infrared** IRAS, MSX, Akari, WISE; **Radio** NVSS, FIRST, PKS, LOFAR, MWO

Classical 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.

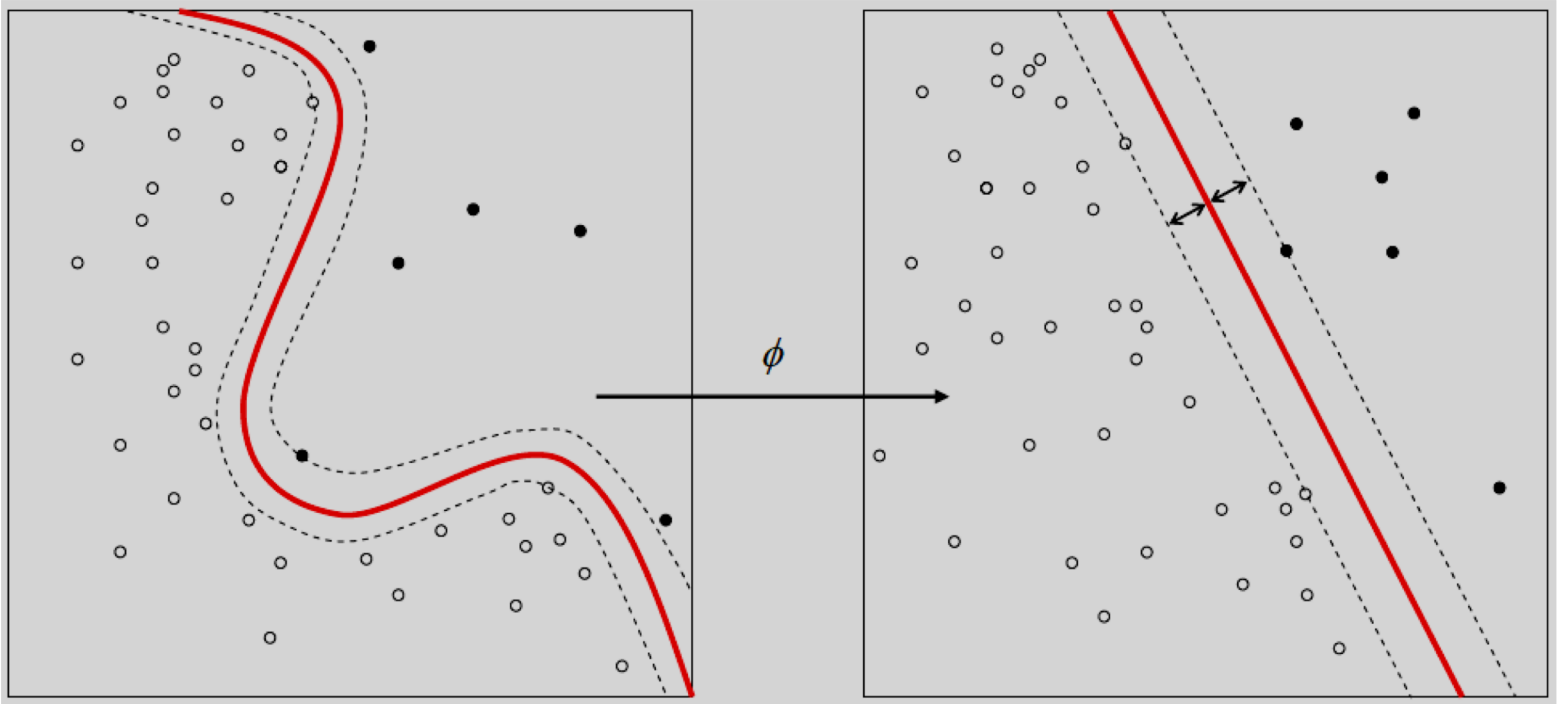
Linear classifiers

Linear discriminant analysis (LDA, 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 separating plane can be interpreted scientifically and used to classify new objects.

Nonparametric linear classifiers that relax the MVN assumption include the **perceptron algorithm** (1950s) that lies at the foundation of **artificial neural networks**, and the **naïve Bayes classifier.**

Support Vector Machines (SVM), developed by Vladimir Vapnik from the 1960-1990s, have emerged as extremely powerful generalizations of LDA and the perceptron. It allows nonlinear surfaces to separate curves in *p*-space and `soft' margins to permit misclassifications. SVMs are very powerful and widely used today.

Support Vector Machines (SVMs), developed by Vladimir 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:

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

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*.

As with KDE, the method encounters problem where both high and low density of data points are present …. we need an adaptive smoother.

Discriminant Adaptive Nearest Neighbor (DANN) is a method where the distance metric (i.e. the standardization of the variables) shrinks when the local density of points is high. The method is related to LDA. The locally distorted distance metric allows classification in the presence of highly-inhomogeneous, elongated and curved structures in *p*-space.

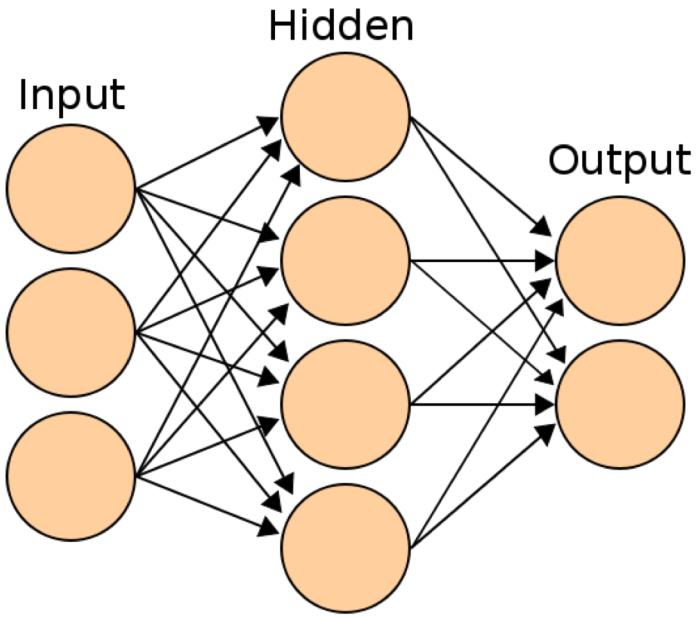
Other methods for difficult classification problems include:

- Machete & Scythe recursive partitionings (Friedman 1994)
- ADAptive MEtric Nearest Neighbor algorithm (Domeniconi et al. 2002)
- Large Margin Nearest Neighbor based on local Mahalanobis distance metrics (Weinberger & Saul 2009)
- Adaptive k-nn classification is related to Bayesian diffusion decision model in neuroscience (Noh et al. 2012)
- …. …. …. [extensive research in data mining techniques]

Artificial 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.

Deep Learning

Recently, superb classification of very difficult problems has been achieved using convolutional neural networks with many hidden layers. Requires very large training sets and very heavy computing. Example:

translate.google.com

English ⟷ Chinese

Deep Learning is transforming modern industries like: social media information propagation, speech recognition, targeted advertising, autonomous vehicles, biometrics & video surveillance, military operations, etc., etc.

Papers in the astronomical literature using Deep Learning are rapidly appearin: i~1/month in 2017, ~1/week in 2018, ~1/day in 2019.

Text: ***Deep Learning***, 2016 I. Goodfellow, Y. Bengio, A. Courville, MIT Press

Final remarks

The word `classification' appeared in ~7000 astronomy papers in 2018 (25%). 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 quantitative methods are increasingly used:

- If no prior knowledge on classes is available, then parametric mixture model or (very uncertain) nonparametric clustering methods
- If prior knowledge is available, then a vast suite of powerful supervised classification methods are available: SVMs, CARTs with boosting & bagging, k-NNs, ANNs

For complex classification problems (e.g. 20 classes in 10-dimensional space with non-MVN structures), parametric models may not be effective while nonparametric methods (CART, k-NN classifiers, ANN) can be successful. Large and reliable training sets are needed for such problems. Deep Learning will allow brilliant new advances in astronomy.