Entry on *Classification* for update volume 2 of the *Encyclopedia of Statistical Sciences*. Wiley, 1996?

The original entry on 'Classification' by Hartigan defines classification as 'the grouping together of similar things'. However, this is only one of its two accepted meanings, and the other, to assign examples to pre-defined classes, is at least as common both in ordinary usage and the specialist literature. Examples include classifying animals ('that's an elephant'), plants, wines, faces, symbols, old furniture (as of a certain school, or as a fake), applicants for credit and stock market opportunities. Whereas a few specialists defined the classes, users take these as given, so the predictive use of classification is far more common.

In the pattern recognition literature [12, 10, 16, 36] the two meanings are called *unsupervised* and *supervised* pattern recognition*, with around 90% of that literature being devoted to supervised problems. This update concentrates on classification in the supervised sense; Kaufman & Rousseeuw [24] is an excellent recent reference on classification in Hartigan's sense. The term *discriminant analysis** is also sometimes used to mean supervised classification, but its original uses (such as Fisher's [13]) were concerned with describing the differences between the groups, not assigning future examples to groups. Supervised pattern recognition has made great progress in the last decade, so this entry can be seen as also updating those on pattern recognition* and discriminant analysis*.

In supervised classification we are given K classes, and the task is to classify further examples by assigning them to one of the K classes. In careful systems two further actions are allowed; to declare the example to have come from none of these classes (an *outlier* decision) or to declare that the classification is 'in doubt' when two or more of the classes are quite plausible. Examples which are not given a firm classification are referred to a second level of classification; perhaps a human expert or to be given some further, expensive, tests. Classifiers are now widely used, for example to read Zip codes on (type-written) envelopes, to alert officials to drugs in airline luggage and to screen cervical smears for abnormal cells. Notice that most of these tasks have no aspect of discrimination; there is no interest in describing the differences between the a's and c's on a type-written envelope, only in reading them. However, there are many tasks for which the ability to explain the classifier is very desirable, a point stressed in the field of *machine learning* [27]. For example, classification systems are beginning to be developed for medical diagnosis, and it is widely believed that these will only be acceptable to medical practitioners if they can 'explain' their reasoning. Our experience has been that even when 'black-box' systems have a very much lower error rate, clients prefer systems which they can understand, for example a simple set of rules.

Edgar Anderson's iris data discussed by Fisher [13] has been very widely used to illustrate classification, and complex 'black-box' classifiers have been built. Yet some very simple rules do almost as well. For example, the three species are almost completely separated on petal area, and Weiss & Kapouleas [43] found the rules

> If Petal length < 3 cm then *I. setosa*.
> If Petal length > 4.9 cm or Petal width > 1.6 cm then *I. virginica*.
> Otherwise *I. versicolor*.

which are easy to interpret, explain the differences between the groups (if not for Fisher's genetical purpose) and perform as well as any other method.

The issues of outliers and missing data are important but not often discussed nor implemented. Ripley [36] does discuss several approaches. In some medical applications

the abnormal examples are of most interest; cancerous tissue on mammograms [40] and abnormal sleep and heart rhythms.

THEORY

Almost all methods of classification can be seen as ways to approximate an optimal classifier, the Bayes rule. Suppose we are given a future example x which actually came from class c. Then if we consider all mis-classifications as equally bad (unit cost) and declare doubt to have cost $0 < d < 1$, the classifier with smallest expected cost finds the class k with the largest posterior probability $p(k \mid x)$, and allocates the example to this class unless all the posterior probabilities are less that $1 - d$, when the example is declared to be 'in doubt' [36]. (This result can be extended quite easily to tasks in which different errors have different costs, but to do so would cloud the main issues here.)

Where do the posterior probabilities come from? Very occasionally they may be specified by a model of the underlying mechanism (one could envisage such examples in population genetics), but normally the probabilities are *learned* from a *training set*, a collection of classified examples. This terminology mimics a common way humans do classification tasks, by comparing the current example with some summary of past experience. Two contrasting points of view have been taken. In what Dawid [9] calls the *diagnostic* paradigm, the posterior probabilities $p(k \mid x)$ are learned directly from the training set. For example, we could take a parametric family $p(k \mid x; \theta)$ and estimate θ by maximizing the conditional likelihood of the observed classes given the observed x's, the approach of *logistic discrimination*. (See LOGISTIC REGRESSION.) The other approach, Dawid's *sampling paradigm*, is to describe the proportions π_k of the classes in the whole population and to describe the examples from class k by a distribution with density $p_k(x)$. Each of these is estimated, and the class k maximizing $p(k \mid x) \propto \pi_k p_k(x)$ is reported. The best-known example of this approach is linear discriminant analysis, in which the class-conditional densities are taken to be multivariate normal with different means but a common covariance matrix, and some recent examples [39, 38, 20] used mixtures of such normal distributions.

This approach, of estimating the posterior probabilities, then acting as if they are the true probabilities, is called the *plug-in* approach, and is often used uncritically. It can result in *over-fitting*, by producing a classifier which works very well on the training set, but performs poorly on future examples. In the parlance of the field, the classifier shows poor *generalization*. This phenomenon has been known for many years, but has become increasingly important as we have learnt how to fit very flexible models. The alternative approaches are all Bayesian in flavour. Adding a prior distribution over θ which discourages extreme posterior probabilities and using the posterior mode of θ leads to an approach known as *regularization*. The approach of *predictive analysis** [1, 17] is more fully Bayesian, and instead of plugging-in just one value $\widehat{\theta}$, averages $p(k \mid x; \theta)$ over the posterior distribution of θ given the training set.

LINEAR METHODS

The best-known method of classification is *linear discriminant analysis*. (See also DISCRIMINANT ANALYSIS.) All the class-conditional densities are taken to be normal with means μ_k and covariance matrix Σ. The plug-in approach chooses the class k which maxi-

mizes

$$L_k = \widehat{\mu}_k^T \widehat{\Sigma}^{-1} x - \tfrac{1}{2}\widehat{\mu}_k^T \widehat{\Sigma}^{-1} \widehat{\mu}_k + \log \pi_k.$$

(If the probabilities π_k are not known, they are estimated by the proportions of the classes in the training set.) The means and Σ are estimated by maximum likelihood, perhaps adjusting for bias by taking divisor $N - K$ for $\widehat{\Sigma}$. If the classes are not assumed to have the same covariance matrix, we obtain quadratic discriminant analysis [26, 36].

The previous paragraph worked within the sampling paradigm. We can also work out the posterior probabilities:

$$\log p(k \mid x) - \log p(1 \mid x) = (\mu_k - \mu_1)^T \Sigma^{-1} x - \tfrac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \tfrac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log \pi_k - \log \pi_1$$

so the log odds between any two classes are linear functions of x. This suggests parametrizing the posterior probabilities directly via

$$p(k \mid x) = \frac{\exp \beta_k^T x + \alpha_k}{\sum_j \exp \beta_j^T x + \alpha_j} \tag{1}$$

which is true for the common-covariance normal model for the classes, but also for several other models and has the direct interpretation in terms of log-odds. This approach is known as *logistic discrimination*. Note that although the common-covariance normal model leads to logistic discrimination, the parameter estimates in the linear and logistic approaches will be different.

FLEXIBLE DISCRIMINANTS

The field of classification has grown enormously since 1985 by exploiting non-linear methods. We now have available a large number of ways to parametrize and fit functions $f\colon \mathbb{R}^m \to \mathbb{R}^p$, some of which we describe later. These can be used to replace linear methods in at least three ways

(a) Logistic discrimination uses a log-linear model (1) for $\log p(k \mid x)$, which can be extended to non-linear models as

$$p(k \mid x; \theta) = \frac{\exp f_k(x; \theta)}{\sum_j \exp f_j(x; \theta)}.$$

The parameters θ are chosen by maximizing the conditional log-likelihood, which is the sum of $-\log p(c \mid x)$ over the training set for the true class c.

(b) For $K = 2$ classes Fisher [13] showed that his linear discriminant could be found by a linear regression of a dummy variable y (zero for class one and one for class two) on x. This suggests replacing the linear regression by a non-linear regression. The extension to $K > 2$ classes is less clear-cut, but one commonly used procedure is to set *targets* which are the indicators of whether that class is true, and fit these targets by a K-variate non-linear regression. Then a future example is classified by which target is closest, or equivalently by which $f_k(x)$ is largest [34, 35]. A variation on this theme is described by Dietterich & Bakiri [11], in which the targets are taken as widely-separated points in a higher-dimensional space.

(c) Breiman & Ihaka [2] showed how Fisher's result could be extended to $K > 2$ classes. One way to perform linear discriminant analysis is to use a linear regression to reduce the m-dimensional input to one in $K - 1$ dimensions, and then to perform LDA in this space which is that spanned by the canonical variates. They suggested replacing the linear regression by a non-linear one, and this has been taken up recently [21, 34, 35, 36].

Many extensions of linear models have been used to form a classifier in one or more of these ways. Perhaps the simplest extensions of linear models are *additive models* [19] in which an arbitrary smooth transformation is allowed for each of the measurements $x_i, i = 1, \ldots, p$. That is,

$$f_k(x) = \alpha_k + \beta_{1k}\phi_1(x_1) + \cdots + \beta_{pk}\phi_p(x_p) \tag{2}$$

and the functions ϕ_j are chosen from a class of smooth functions, for example cubic smoothing splines. (Such models were also derived within the framework of additive splines; [41, 42].)

There are two natural ways to extend (2). One is to allow interactions, that is terms of the form $\phi_i(x_i)\phi_j(x_j)$. These are considered within the spline framework [41] (see also SPLINE FUNCTIONS) and as Friedman's [14] *multivariate adaptive regression splines*. The other way is to replace individual measurements with linear combinations, to obtain

$$f_k(x) = \alpha_k + \sum_{j=1}^{r} \beta_{jk}\phi_j(a_j^T x + b_j). \tag{3}$$

This is *projection pursuit regression* [15].

One special case of (3) is particularly important. This is when each ϕ_j is the *logistic* function $\ell(x) = e^x/(1 + e^x)$, and (3) is the simplest and by far the most popular form of *neural network*, sometimes called a *feed-forward neural network*, a *backpropagation network* or a *multi-layer perceptron*. The form (3) is referred to as a network with linear output units; when it is used in a log-linear model as in approach (a), it is called a *softmax* [4] network. Another form of log-linear model is more commonly used for classification by neural networks. This has

$$p(k \mid x; \theta) = \ell\Big(\alpha_k + \sum_{j=1}^{r} \beta_{jk}\ell(\sum_i a_{ij}x_i + b_j)\Big). \tag{4}$$

and corresponds to linked non-linear logistic regressions for each class. This would be appropriate in medical diagnosis, where we wish to know the probability that each of K diseases is present, but not for general classification problems where softmax should be used.

Another class of flexible functions, *radial basis functions* (or RBFs), has been borrowed from interpolation theory [18, 30] and taken up by the neural network community [5, 28, 29]. These are of the form

$$f_k(x) = \alpha_k + \sum_j \beta_{jk} G(\|x - x_j\|)$$

for some points x_j and a function G which is often taken to be $\exp -\frac{1}{2}(x/\sigma)^2$ but can also be unbounded like $\sqrt{c^2 + r^2}$.

These classes of functions really are very flexible. Each of projection pursuit regression, neural networks and radial basis functions can approximate any function $f: \mathbb{R}^m \to \mathbb{R}^p$ arbitrarily well in a number of senses, including uniform convergence on compact sets and L_2 convergence. Thus although they have parameters, they are non-parametric in spirit, and can easily over-fit the training set.

The issues of using these families of non-linear functions well are many and complex, and the subject of on-going research. One crucial issue is how complex the family of functions should be chosen to be; it has to be complex enough to model the true posterior probabilities yet there has to be some restraint to avoid over-fitting the particular training set at hand. Unfortunately, the literature shows that there is ample scope to misuse the new-found power to build very complex classifiers.

NON-PARAMETRIC APPROACHES

The term non-parametric is not very useful in classification, since there is no essential difference between methods with a very large number of parameters (such as neural networks and general mixtures of normal distributions) and those with none (kernel density estimation and non-parametric regression).

The obvious non-parametric approach within the sampling paradigm is to use kernel density estimates of $p_k(x)$. This has not proved at all successful. The number of measurements m is often large, and we need to decide which $\pi_k p_k(x)$ is largest, which will often mean looking at x which is in the tails of all the distributions.

Non-parametric approaches within the diagnostic paradigm have been much more successful. One of the simplest methods of classification, that of k-nearest neighbours (see also NEAR-NEIGHBOUR ESTIMATION) is also one of the most successful. Given a new example, we find the k nearest examples in the training set and take a vote amongst their classes to classify the new example. (Ties are usually broken at random.) This method has several potential disadvantages. We have to choose the measure of dissimilarity used to select the nearest examples. We have to find those examples (although several methods to find them quickly are known [8, 36]), and we have to store the examples. Many methods have been proposed to replace the training set by a smaller set of representatives of each class; one of the most promising is Kohonen's *Learning Vector Quantization* [25, 36].

Classification trees (see also RECURSIVE PARTITIONING) provide another example of a successful non-parametric method within the diagnostic paradigm. These partition the space of measurements, and assign estimates of the posterior probabilities to each cell of the partition. (These estimates are usually the frequencies of the classes amongst the examples from the training set which fall into the cell.) The partitioning is done recursively, one measurement at a time. In a real example [33] the first split is on the minimum temperature, and then those sites with minimum temperature 16° or greater are split on evaporation, and those with high evaporation are further divided by the maximum temperature. This binary tree of splits leads to the name.

Classification trees are best known within statistics from the CART book [3] and program, but they were developed independently at about the same time both in machine learning [31] and engineering [37]. It has proved easy to suggest myriads of methods of growing a tree, but difficult to find reliable ways to select the size of the tree. The most effective suggestions have been to grow quite a large tree, and then to prune it to optimize some

measure of performance [3, 32, 36], but it is not uncommon to be left with a large range of uncertainty over the size of the tree unless the training set is large (several thousands).

Further Reading

Ripley [36] provides the most comprehensive overview of supervised pattern recognition, as well of neural networks from a statistical viewpoint. (The articles [33, 34, 6] discuss other statistical aspects of neural networks.) McLachlan [26] covers the classical statistical approaches to supervised classification. The books by Haykin [22] and Hertz, Krogh & Palmer [23] provide comprehensive introductions to the field of neural networks, and Quinlan [32] is perhaps the most accessible account of classification methods developed by the machine learning community. The edited volumes [27, 7, 44] provide a number of excellent introductory and review articles spanning many approaches.

Bibliography

Bishop, C. (1995) *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford. [An introduction to neural networks slanted towards classification.]

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Monterey, CA.

Cheng, B, and Titterington, D. M. (1994) Neural networks: a review from a statistical perspective (with discussion). *Statistical Science* **9**, 2–54.

Cherkassky, V., Friedman, J. H. and Wechsler, H. (eds) (1994) *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.* Springer, Berlin.

Dasarathy, B. V. (ed.) (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA. [A collection of key articles.]

Devijver, P. A. and Kittler, J. V. (1982) *Pattern Recognition. A Statistical Approach.* Prentice-Hall, Englewood Cliffs, NJ. [Mainly on k-nearest neighbour methods.]

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis.* Wiley, New York. [An updated version by Duda, Hart and Stork is planned.]

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition.* Second edition. Academic Press, Boston.

Hand, D. J. (1982) *Discrimination and Classification.* Wiley, Chichester.

Haykin, S. (1994) *Neural Networks. A Comprehensive Foundation.* Macmillan College Publishing, New York. [This really is comprehensive and has over a thousand references.]

Hertz, J., Krogh, A. and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation.* Addison-Wesley, Redwood City, CA.

McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York. [Encyclopedic on the statistical literature, with over a thousand references.]

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (eds) (1994) *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, New York. [Introductory survey articles plus a serious comparative study.]

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.

Ripley, B. D. (1993) Statistical aspects of neural networks. In *Networks and Chaos – Statistical and Probabilistic Aspects* eds O. E. Barndorff-Nielsen, J. L. Jensen and W. S. Kendall, pp. 40–123. Chapman & Hall, London.

Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. R. Statist. Soc. B* **56**, 409–456.

Ripley, B. D. (1995) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge. [Comprehensive, with over a thousand references.]

REFERENCES

[1] Aitchison, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis.* Cambridge University Press, Cambridge.

[2] Breiman, L. and Ihaka, R. (1984) Nonlinear discriminant analysis via ACE and scaling. Technical Report 40, Dept. of Statistics, University of California, Berkeley.

[3] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Monterey, CA.

[4] Bridle, J. S. (1990) Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures and Applications* eds F. Fogleman Soulié and J. Hérault, pp. 227–236. Springer, Berlin.

[5] Broomhead, D. S. and Lowe, D. (1988) Multivariable functional interpolation and adaptive networks. *Complex Systems* **2**, 321–355.

[6] Cheng, B, and Titterington, D. M. (1994) Neural networks: a review from a statistical perspective (with discussion). *Statistical Science* **9**, 2–54.

[7] Cherkassky, V., Friedman, J. H. and Wechsler, H. (eds) (1994) *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.* Springer, Berlin.

[8] Dasarathy, B. V. (ed.) (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA.

[9] Dawid, A. P. (1976) Properties of diagnostic data distributions. *Biometrics* **32**, 647–658.

[10] Devijver, P. A. and Kittler, J. V. (1982) *Pattern Recognition. A Statistical Approach.* Prentice-Hall, Englewood Cliffs, NJ.

[11] Dietterich, T. G. and Bakiri, G. (1995) Error-correcting output codes: a general method for improving multiclass inductive learning programs. In [44], pp. 395–407.

[12] Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis.* Wiley, New York.

[13] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

[14] Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–141.

[15] Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817–823.

[16] Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition.* Second edition. Academic Press, Boston.

[17] Geisser, S. (1993) *Predictive Inference: An Introduction.* Chapman & Hall, New York.

[18] Hardy, R. L. (1971) Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **76**, 1906–1915.

[19] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models.* Chapman & Hall: London.

[20] Hastie, T. and Tibshirani, R. (1995) Discriminant analysis by Gaussian mixture. *J. R. Statist. Soc. B* (to appear).

[21] Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89**, 1255–1270.

[22] Haykin, S. (1994) *Neural Networks. A Comprehensive Foundation.* Macmillan College Publishing, New York.

[23] Hertz, J., Krogh, A. and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation.* Addison-Wesley, Redwood City, CA.

[24] Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data. An Introduction to Cluster Analysis.* Wiley, New York.

[25] Kohonen, T. (1995) *Self-Organizing Maps.* Springer, Berlin.

[26] McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

[27] Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (eds) (1994) *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, New York.

[28] Moody, J. and Darken, C. J. (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation* **1**, 281–294.

[29] Poggio, T. and Girosi, F. (1990) Networks for approximation and learning. *Proc. IEEE* **78**, 1481–1497.

[30] Powell, M. J. D. (1987) Radial basis functions for multivariable interpolation: a review. In *Algorithms for Approximation* eds J. C. Mason and M. G. Cox, pp. 143–167. Clarendon Press, Oxford.

[31] Quinlan, J. R. (1983) Learning efficient classification procedures and their application to chess end-games. In *Machine Learning* eds R. S. Michalski, J. G. Carbonell and T. M. Mitchell, pp. 463–482. Tioga, Palo Alto, CA.

[32] Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.

[33] Ripley, B. D. (1993) Statistical aspects of neural networks.In *Networks and Chaos – Statistical and Probabilistic Aspects* eds O. E. Barndorff-Nielsen, J. L. Jensen and W. S. Kendall, pp. 40–123. Chapman & Hall, London.

[34] Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. R. Statist. Soc. B* **56**, 409–456.

[35] Ripley, B. D. (1994) Neural networks and flexible regression and discrimination. *Adv. Appl. Statist.* **2**, pp. 39–57.

[36] Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge, Cambridge University Press.

[37] Sethi, I. K. and Sarvarayudu, G. P. R. (1982) Hierarchical classifier design using mutual information. *IEEE Trans. PAMI* **4**, 441–445.

[38] Streit, R. L. and Luginbuhl, T. E. (1994) Maximum likelihood training of probabilistic neural networks. *IEEE Trans. Neural Networks* **5**, 764–783.

[39] Taxt, T., Hjort, N. L. and Eikvil, L. (1991) Recognition of handwritten symbols. *Patt. Recgn. Lett.* **12**, 731–737.

[40] Tarassenko, L., Hayton, P., Cerneaz, N. and Brady, M. (1995) Novelty detection for the identification of masses in mammograms. In *Proc. Fourth Int. IEE Conf. Artificial Neural Networks (Cambridge, 1995).* IEE Press.

[41] Wahba, G. (1990) *Spline Models for Observational Data.* SIAM, Philadelphia, PA.

[42] Wahba, G., Gu, C., Wang, Y. and Chappell, R. (1995) Soft classification a.k.a. risk estimation via penalized log likelihood and smoothing spline analysis of variance. In [44] pp. 331–359.

[43] Weiss, S.M. amd Kapouleas, I. (1989) An empirical comparison of pattern recognition, neural nets and machine learning classification methods. *Proc. 11th International Joint Conference on Artificial Intelligence, Detroit, 1989*, pp. 781–787.

[44] Wolpert, D. H. (ed.) (1995) *The Mathematics of Generalization.* Addison-Wesley, Reading, MA