# CLASSIFICATION: ASTRONOMICAL AND MATHEMATICAL OVERVIEW

F. MURTAGH[†]
*Space Telescope - European Coordinating Facility*
*European Southern Observatory*
*Karl-Schwarzschild-Straße 2*
*D-85748 Garching, Germany*

SUMMARY. A short overview is presented of a number of different discriminant analysis methods. An 'uncertainty principle' is presented in regard to the issue of user choice of appropriate method. The discriminant analysis methods described are then used in the important problem of feature selection. A hand-classified set of HST Guide Star plate data is used, with star/galaxy/fault classes.

## 1. Introduction

In classification, prior knowledge of class membership may not be available, in which case cluster analysis or unsupervised classification methods can be used. Where class labels are available for the objects being studied, discriminant analysis or supervised classification methods are used. Often methods of the latter type are more likely to be in question when 'classification' is spoken of.

Discriminant analysis methods are often subdivided into parametric and non-parametric methods. Parametric methods seek generality by assuming that the objects follow some statistical model. For example, the measurements made on objects in a particular class may be assumed to be distributed as a multivariate Gaussian distribution. This, then, implies that a first requirement is that we estimate the relevant model parameters, such as the mean and covariances.

Non-parametric methods seek to be data-driven. A training set is used to train the classifier, i.e. to arrive at estimates for its parameters (e.g. the weights in a multilayer perceptron). Right away we see that a good training set is a critical requirement for good performance. The latter is referred to as generalisation or application, when the trained classifier is used to produce class memberships for objects which were not formerly used for training. These test objects can be used for assessment of performance. Resubstitution, using the training set, gives an over-optimistic estimate of the error rate. Since dividing one's data into a training and a test set lessens the available data for training the classifier, a popular validation strategy is referred to as leaving-one-out. Given $n$ objects, each object is considered in turn as a test object, with the training of the classifier carried out on the remaining $n$ - 1 objects.

Although these approaches to supervised classification are quite varied, nonetheless there are

---

[†] Affiliated to Astrophysics Division, Space Science Department, European Space Agency



*H. T. MacGillivray et al. (eds.), Astronomy from Wide-Field Imaging,* 227–233.

many points of overlap. Some methods can be arrived at both by parametric and by non-parametric considerations. Various relationships have been studied between the different methods. A very short presentation of some methods, beginning in the next section, will be followed by an interesting characterization of the choice of parametric versus non-parametric methods. Following this, we will use the methods described to present a few results on the issue of best features for star/galaxy separation.

## 2. A Geometric Formulation

A geometric formulation of the discrimination problem is based on an $m$-dimensional feature space (i.e. each object is associated with its set of $m$ features) mapped onto a new, Euclidean space which takes account of the classes. In section 6, it will be shown that this works well, in particular if the features are well chosen. Following a presentation of this method in this section, the next section will show how the same separation surface between classes may be derived by a probabilistic and Bayesian argument. A general treatment of the mathematical formalism required for these analyses may be found in Duda & Hart (1973).

Notation: Object vectors are row vectors of $X = \{x_{ij} : i \in I, j \in J\}$ for a finite set of objects, $I$, and a finite set of variables, $J$. The grand mean is $g_j = 1/n \sum_{i \in I} x_{ij}$ for all $j \in J$ where $n$ is the cardinality of the object set, $I$. The mean of group $p$ has $j$th coordinate: $p_j = 1/n_p \sum_{i \in p} x_{ij}$. Let a partition of $I$ be denoted $P$, with $p \in P$.

$T$ is the total variance-covariance matrix; $B$ is the between-groups variance-covariance matrix; and $W$ is the within-groups variance-covariance matrix. We have that: $T = W + B$.

$$\mathbf{T}: \quad t_{jk} = \frac{1}{n} \sum_{i \in I} (x_{ij} - g_j)(x_{ik} - g_k)$$

$$\mathbf{W}: \quad w_{jk} = \frac{1}{n} \sum_{p \in P} \sum_{i \in p} (x_{ij} - p_j)(x_{ik} - p_k)$$

$$\mathbf{B}: \quad b_{jk} = \sum_{p \in P} \frac{n_p}{n} (p_j - g_j)(p_k - g_k)$$

Multiple discriminant analysis (or canonical discriminant analysis, or discriminant factor analysis) seeks an axis (and, subsequently, orthogonal axes) $\mathbf{u}$ such that the spread of group means is large, while restraining the spread within groups to be small. This is mathematically expressed as : maximize $\mathbf{u}'B\mathbf{u}$ and minimize $\mathbf{u}'W\mathbf{u}$, simultaneously (where $'$ denotes transpose).

This discriminant analysis method can be characterised as principal components analysis of the group means, in the Mahalanobis or $T^{-1}$-metric. The distance of an object to a group is $(x_{ij} - p_j)'T^{-1}(x_{ij} - p_j)$. The Mahalanobis distance has the effect of adjusting for variations in spread of the cloud of object-points.

In the two-group case, this is Fisher's widely-used linear discriminant analysis. The separation surface between two groups is derived from that part of the space which is closer to one group than to the other, with respect to the Mahalanobis metric. Hence, in this space, a separation surface is a hyperplane. An identical algorithm can be derived from a parametric perspective, as will now be shown.

## 3. A Parametric Formulation

Given members of a group, $p$, we can sample the values of their variables. Thus we can obtain estimates of the probability of having certain variable values, given membership in the group $P(\boldsymbol{x} \mid p)$. To achieve generality, the variable values associated with a group are often modelled. Assuming that the group is approximated by a multivariate Gaussian, we have:

$$P(\boldsymbol{x} \mid p) = (2\pi)^{-\frac{n}{2}} \mid V_p \mid^{\frac{1}{2}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{p})' V_p^{-1}(\boldsymbol{x} - \boldsymbol{p})) \tag{1}$$

where we have written $p$ for the group (considered as a set), and also $\boldsymbol{p}$ for the group's mean vector, $V_p$ is the variance-covariance matrix of the group, and $\mid . \mid$ is the determinant. We must estimate the mean vector $\boldsymbol{p}$ and the matrix $V_p$, i.e. the model parameters.

In practice, we need to know $P(p \mid \boldsymbol{x})$. To determine this, use is made of Bayes' theorem:

$$P(p \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid p)P(p)}{\sum_{all\ q} (P(\boldsymbol{x} \mid q)P(q))} \tag{2}$$

By calculating the right hand side for two groups, $p_1$ and $p_2$, on the basis of a given set of values for the vector $\boldsymbol{x}$, we can choose $p_1$ in preference to $p_2$ if $P(p_1 \mid \boldsymbol{x}) > P(p_2 \mid \boldsymbol{x})$. This is the Bayes minimum risk criterion, which can be solved using equs. 1 and 2.

If variances and covariances $V_{p1}$ and $V_{p2}$ are considered to be the same for two groups, $p_1$ and $p_2$, then the decision rule is a linear one. Take $V_p$ now as the population variance-covariance matrix, denoted $T$ in section 2 above. If the groups' prior probabilities, $P(p_1)$ and $P(p_2)$, are the same, then this decision rule is *identical* to the Fisher linear discriminant analysis one.

## 4. K-Nearest Neighbours Discriminant Analysis

K-nearest neighbours discriminant analysis is used below in section 6. The decision surfaces between classes are piecewise linear, so the net effect is that of a nonlinear classification method. It shares this property with the multilayer perceptron and with classification (or decision) trees. In the next section, section 5, we will raise the question as to the conditions under which nonlinear (and non-parametric) methods of these types might be preferred (or otherwise) to methods based on statistical modelling.

In k-nearest neighbours, a non-parametric estimation of $P(p \mid \boldsymbol{x})$ is obtained as follows. Consider volume $v$ around $\boldsymbol{x}$. Suppose $k$ cases are 'captured' in $v$. Suppose further that $k_p$ of these are associated with (or labelled) group $p$.

The approximate joint density function of $\boldsymbol{x}$ and $p$ is $P(\boldsymbol{x}, p) = (k_p / n)/v$, where there are $n$ cases in total.

Now, by definition $P(p \mid \boldsymbol{x})\ P(\boldsymbol{x}) = P(\boldsymbol{x}, p)$. Hence,

$$P(p \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, p)}{\sum_{all\ q} P(\boldsymbol{x}, q)}$$

https://doi.org/10.1017/S0074180900047379 Published online by Cambridge University Press

$$= \frac{(k_p / n) / V}{\sum_{all\ q} (k_q / n) / V} = \frac{k_p}{k}$$

This is the fraction of cases labelled $p$ in the volume around $\boldsymbol{x}$. Therefore $\boldsymbol{x}$ is labelled group $p$ if the majority of its $k$ nearest neighbours are labelled this.

Using diverse datasets, we have found the multilayer perceptron to yield results which are very similar to k-nearest neighbours (Murtagh 1992). This is not surprising since the transfer functions in the neural net classifier can be regarded as imposing a two-way split on their inputs (hence a linear split on each neuron's scalar throughput). The compound effect of linear splits of the data, occasioned at each node of the multilayer perceptron, is reminiscent of what is done by the k-nearest neighbours approach (Lippmann 1987). Turning to the parametric perspective, error bounds for the k-nearest neighbour method can be expressed in terms of the Bayesian error rate (Duda & Hart 1973). Two recent papers which experimentally compare results obtained by k-nearest neighbours, multilayer perceptrons, and classification (or decision) trees are Ripley (1993a, 1993b).

## 5. The Bias/Variance Dilemma

The methods sketched out in previous sections have been described in terms of two major alternative approaches: either statistical modelling of the classes is attempted, or else a flexible and data-sensitive method is sought. Both general approaches have advantages and disadvantages. Since every data set is different, it is difficult to favour one method over another in all circumstances. When skilfully and carefully used, methods based on quite different assumptions can give rise to roughly comparable results. In this section we will briefly present an aspect of discrimination methods which clarifies the polarity between parametric and non-parametric approaches.

Discriminant analysis provides us with a mapping, which takes given input vectors $\boldsymbol{x}$, and produces an output value (or vector, depending on the problem), $f(\boldsymbol{x})$. We would like this output to be as close as possible, on average, to a desired $y$.

Without undue loss of generality, consider a squared discrepancy estimate of error, $E[(y - f(\boldsymbol{x}))^2 \mid \boldsymbol{x}]$. Here, $E$ is the expected value, and $y$ is the function we wish to fit to our function of $\boldsymbol{x}$. A prediction in regression would then be $E(y \mid \boldsymbol{x})$. In classification, we can define $y = 1$ if $\boldsymbol{x} \in$ group 1, and $y = 0$ otherwise.

The following relationship can be established (see German et al. 1992):

$$E[(y - f(\boldsymbol{x}))^2 \mid \boldsymbol{x}] = E[(y - E[y \mid \boldsymbol{x}])^2 \mid \boldsymbol{x}] + (E[y \mid \boldsymbol{x}] - f(\boldsymbol{x}))^2$$

The first term on the right hand side is the variance of $y$ given $\boldsymbol{x}$; it is the scatter of predicted $y$ values about the function $y$. The second term on the right hand side is the effectiveness of $f$ as a predictor of $y$, i.e. the prediction minus the given input function.

This equation decomposes estimation error into variance plus bias terms. German et al. (1992) propose it as a fundamental uncertainty principle of statistical inference, shared by problems such as regression and classification, and shared equally by non-parametric statistical and neural network methods. For a given error, one may improve on the bias, but at the expense of the variance, or vice versa.

Model-based approaches can suffer from high bias, i.e. incorrect predictions due to over-straightjacketing the data. Non-parametric methods can suffer from high variance, i.e. "letting the data speak" implies high sensitivity to the data, which in turn means that a very large number of training cases are needed in order to pin down good predictions.

Modelling may be algorithmically difficult, but is not conceptually so. On the other hand, for non-model based approaches, the choice of training data may be problematic. The dimensionality of the space of variables affects model-based approaches in regard to computational aspects of the optimization of model fits. On the other hand, increasing spatial dimensionality can hugely accentuate the problem of acceptable behaviour of non-parametric methods on the training set, coupled with potentially very bad behaviour on test sets.

## 6. Feature Selection

The appropriate choice of features is an important issue. Trivially, we can note that no amount of sophistication in the discriminant analysis method used can allow very badly chosen features to discriminate between classes. Conversely, we show below that even linear separation between classes performs well when the classifier is provided with reasonably good features.

The data used comprised 628 hand-classified objects from Guide Star Scan plates: 549 stars, 48 galaxies, 27 faults and 4 weak faults. For convenience, the latter two groups were merged. These objects were for the most part of intermediate brightness — not near the noise limit, nor very large.

Four sets of features were used:

1) 'traditional' variables: log total luminosity; ratio of peak to total luminosity; surface brightness for two different thresholds; two spike values; ellipticity; area ratio; offset from peak to centroid;
2) Texture features (Haralick et al. 1973), averaged over 4 angles, for an 8-fold quantization of the grey levels: angular second moment; contrast; correlation; variance; inverse difference moment; sum average; sum variance; sum entropy; entropy; and others. Motivation for the use of texture measures may be found in various papers of Malagnini (e.g. 1983);
3) spreads (max minus min) of the foregoing texture features. The motivation is that a range of variation might be a more sensitive discriminatory measure;
4) a new texture-like set of 12 features (not further explained here).

The method initially used was as follows. Perform multiple discriminant analysis on data; determine projections of group centres in the new coordinate system; for each case in turn (using coordinates in the new coordinate system), determine its squared Euclidean distance to each group, and hence determine the closest group centre; check if this was the correct group, and build a contingency table to summarize all results.

Results for the 'Traditional' or first set of features are given in Table 1, while those obtained using all four sets of variables together (46 variables altogether) are given in Table 2.

**Table 1.** Traditional features

| Assignment to class: | 1 | 2 | 3 |
|---|---|---|---|
| Known class 1 (stars) | 530 | 12 | 7 |
| Known class 2 (galaxies) | 11 | 22 | 15 |
| Known class 3 (faults) | 4 | 18 | 9 |

Percentage correct (i.e. (530+22+9)/628) = 89.3%
Star purity (i.e. 530/(530+12+7)) = 96.5%

**Table 2.** All four sets of features

| Assignment to class | 1 | 2 | 3 |
|---|---|---|---|
| Known class 1 (stars) | 537 | 10 | 2 |
| Known class 2 (galaxies) | 9 | 27 | 12 |
| Known class 3 (faults) | 3 | 15 | 13 |

Percentage correct (i.e. (537+27+13)/628) = 91.9%
Purity of stars (i.e. 537/(537+10+2)) = 97.8%

We conclude that even linear methods give good results, when the right features are used.

We next ask if non-linear methods can squeeze more performance out of the data, e.g. with K-nearest neighbours, k = 9, using 'traditional' or first set of features. Table 3 gives cross-validation results, using leave-one-out estimates. Proximity of cases to their nearest neighbours used the Mahalanobis distance, proportional priors were used.

**Table 3.** Traditional features with K-nearest neighbours

| Assignment to class: | 1 | 2 | 3 |
|---|---|---|---|
| Known class 1 (stars) | 544 | 4 | 1 |
| Known class 2 (galaxies) | 13 | 32 | 3 |
| Known class 3 (faults) | 14 | 4 | 13 |

Percentage correct (i.e. (544+32+13)/628) = 93.8%
Purity of stars (i.e. 544/(544+4+1)) = 99.1%

By using texture-related measures, we probably could have achieved even superior results. However, before doing so we would prefer to re-check the correctness of the hand-classified data! The greater the number of features used, the more accentuated the problem of high-variance estimates (discussed in section 5) becomes.

We plan next to assess which features are of most use, among the rather large set studied. Having done this, we will look at the possibilities for unsupervised classification, or clustering.

## Acknowledgements

Section 6 benefitted from many discussions with B. Lasker. The choice of data, and the visual classification used, also resulted from this work.

## References

Duda, R.O. and Hart, P.E., 1973. 'Pattern Classification and Scene Analysis', Wiley, New York.

Geman, S., Bienenstock, E. and Doursat, R., 1992. 'Neural networks and the bias/variance dilemma', *Neural Computation*, **4**, 1.

Haralick, R.M., Shanmugam, K and Dinstein, I., 1973. 'Textural features for image classification', IEEE Transactions on Systems, Man and Cybernetics, SMC-3, p. 610.

Lippmann, R.P., 1987. 'An introduction to computing with neural nets', *IEEE ASSP Magazine*, April, 4.

Malagnini, M.L., 1983. 'A classification algorithm for star/galaxy counts', *Proc. Statistical Methods in Astronomy*, ESA SP-201, 69.

Murtagh, F., 1992. 'The multilayer perceptron for discriminant analysis: two examples', in Analyzing and Modelling Data and Knowledge, ed. M. Schader, Springer-Verlag, Berlin, p. 305.

Ripley, B.D., 1993a. '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, Chapman and Hall, London.

Ripley, B.D., 1993b. 'Neural networks and related methods for classification', manuscript.