Supervised Classification

In this segment, supervised classification techniques will be studied.

Similarities with unsupervised classification

- ✓ The population has q groups.
- ✓ Variables $X_1, \ldots, X_p$ can be obtained for any individual.

Differences with unsupervised classification

- ✓ The number and identity of the groups are known.
- ✓ The group Y to which each individual belongs is also known.

Ultimate objective: To find a model or algorithm to classify new objects/individuals in the right group, i.e., predict Y from $X_1, \ldots, X_p$.

Examples of Supervised Classification (1–2)

- ✓ An archaeologist wants to determine whether human remains are those of a man or those of a woman.
- ✓ Based on a credit risk analysis, should a loan be granted to an individual or not?
- ✓ Revenu Québec wishes to identify tax returns that deserve to be examined in greater depth (fraud detection).
- ✓ Automatic recognition of letters and numbers of handwritten postal codes.
- ✓ An insurance company wants to predict which policyholders are good and bad drivers and / or whether or not they will have claims.

Examples of Supervised Classification (2–2)

And several other great classics ...

- ✓ Identification of new potential customers
- ✓ Recommendation systems: the ad you see while browsing the web is often not chosen at random!
- ✓ Spam filtering
- ✓ Prediction of winners at sporting events
- ✓ Automated text translation
- ✓ Image recognition

General Approach to Supervised Classification

1. Select a certain number of individuals for whom you know to which group they belong.
2. Measure p characteristics $X_1, \ldots, X_p$ on these individuals.
3. Divide the data set in two parts:
   - a "training" data set for model construction;
   - a "testing" data set for model validation.
4. Develop a model or an algorithm to classify the individuals from the training data set in the best possible way.
5. Evaluate the model or algorithm on the validation data set.
6. Repeat Steps 3–5 with other models or algorithms and choose the best option.

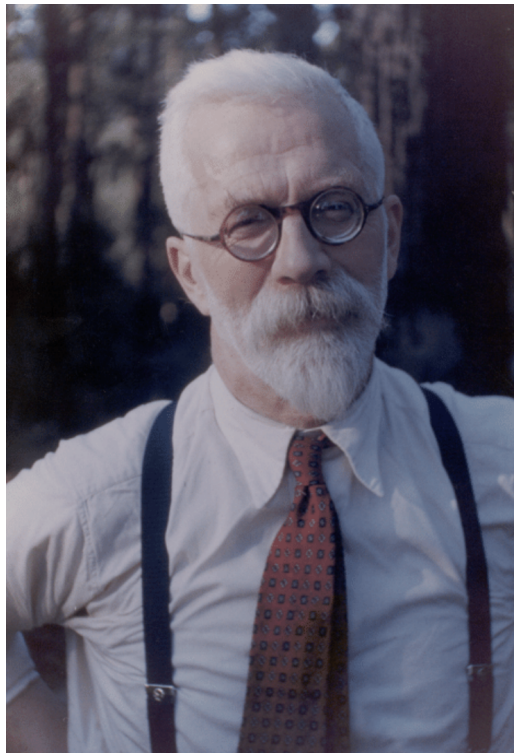
Methods of Supervised Classification

Various methods are available:

- ✓ discriminant analysis;
- ✓ classification trees;
- ✓ logistic regression and related methods, e.g., GLM, GAM, GLMM, GAMM, etc.;
- ✓ support vector machines;
- ✓ neural networks.

Discriminant Analysis

Sir Ronald A. Fisher introduced discriminant analysis in the following article:



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

I. Discriminant functions

When two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (*a*) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (*b*) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

II. Arithmetical procedure

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four

R.A. Fisher (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, vol. 7, pp. 179–188.

Notation

Let $\mathbf{X} = (X_{ij})$ be an $n \times p$ matrix, where

- ✓ n is the sample size,
- ✓ p is the number of variables,
- ✓ for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$,

$$X_{ij} = \text{value of the } j\text{th variable for the } i\text{th individual.}$$

For each group $k \in \{1, \ldots, q\}$, let also

- ✓ I_k be the set of individuals in group k,
- ✓ $n_k = |I_k|$ be the size of I_k,

so that $n_1 + \cdots + n_q = n$.

Discriminant Analysis Score

The observations are assumed to be vectors in $\mathbb{R}^p$.

To classify an observation $(X_1, \ldots, X_p)$, one must partition $\mathbb{R}^p$ into q subsets in such a way that each subset corresponds to a unique group.

Fisher's strategy:

- ✓ Reduce the dimension from p to 1 by computing a score

$$f(X_1, \ldots, X_p) \in \mathbb{R},$$

  for each observation.

- ✓ Use this score to determine the groups, i.e., partition $\mathbb{R}$.

Discriminant Function

Fisher's score is a linear combinaison of the variables, viz.

$$f(X_1, \ldots, X_p) = \mathbf{a}^\top \mathbf{X} + b = a_1 X_1 + \cdots + a_p X_p + b.$$

This function will be used to deduce q intervals $\mathcal{I}_1, \ldots, \mathcal{I}_q$ corresponding to the various groups.

Without loss of generality, one can choose

$$-b = a_1 \bar{X}_1 + \cdots + a_p \bar{X}_p = \mathbf{a}^\top \bar{\mathbf{X}},$$

which centers the variables by removing the mean vector, viz.

$$\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^\top.$$

It only remains to choose the vector $\mathbf{a} = (a_1, \ldots, a_p)$.

Choice of the Best Score

The intuition behind the choice of the vector $\mathbf{a}$ is to ensure that the scores are

- ✓ as homogeneous as possible within a group;
- ✓ as heterogeneous as possible between groups.

Recall that for any (column) vector $\mathbf{a} \in \mathbb{R}^p$, one has

$$\text{var}\{\mathbf{a}^\top (X_1, \ldots, X_p)^\top\} = \mathbf{a}^\top \text{var}\{(X_1, \ldots, X_p)^\top\}\mathbf{a},$$

which can be estimated from a sample of n observations using

$$\mathbf{a}^\top \mathbf{S}\mathbf{a}/n.$$

Variance Decomposition

Discriminant analysis is based on the fact that

$$\mathbf{S} = \mathbf{W} + \mathbf{B},$$

where

$$\mathbf{W} = \text{within-group variance matrix}$$

and

$$\mathbf{B} = \text{between-group variance matrix.}$$

We will now see how this is done.

A Bit of Notation

The mean of variable j based on all individuals in the sample is

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}.$$

The mean of variable j based on all individuals in group k is

$$\bar{X}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} X_{ij}.$$

The total sum of squares is

$$s_{jj'} = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right) \left(X_{ij'} - \bar{X}_{j'}\right).$$

Decomposition of $s_{jj'}$

One could derive from $\mathbf{S} = (s_{jj'})$ an estimation of $\text{cov}(X_j, X_{j'})$ if all the observations came from the same group.

Now each entry in $\mathbf{S}$ can be decomposed as follows:

$$s_{jj'} = w_{jj'} + b_{jj'},$$

where

$$w_{jj'} = \sum_{k=1}^{q} \sum_{i \in I_k} (X_{ij} - \bar{X}_{kj})(X_{ij'} - \bar{X}_{kj'}),$$

$$b_{jj'} = \sum_{k=1}^{q} n_k (\bar{X}_{kj} - \bar{X}_j)(\bar{X}_{kj'} - \bar{X}_{j'}).$$

Sketch of Proof

1. Set
$$X_{ij} - \bar{X}_j = X_{ij} - \bar{X}_{kj} + \bar{X}_{kj} - \bar{X}_j,$$
in the definition of $s_{jj'}$, and same for $X_{ij'}$.

2. Develop the products.

3. Replace $\sum_{i=1}^{n}$ by $\sum_{k=1}^{q}\sum_{i \in I_k}$.

4. Make the appropriate simplifications.

Consequence of the decomposition of $s_{jj'}$

One finds

$$\widehat{\text{var}}\{\mathbf{a}^\top (X_1, \ldots, X_p)^\top\} = \frac{1}{n}\,\mathbf{a}^\top \mathbf{S}\mathbf{a} = \frac{1}{n}\left(\mathbf{a}^\top \mathbf{W}\mathbf{a} + \mathbf{a}^\top \mathbf{B}\mathbf{a}\right).$$

Recall that we want to choose the vector $\mathbf{a}$ in such a way that the scores separate the groups as easily as possible.

In other words, we want scores that are as similar as possible within each group and as different as possible between groups.

Therefore, we choose $\mathbf{a} \in \mathbb{R}^p$ in order to maximize

$$\frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \quad \text{or} \quad \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{S}\mathbf{a}} = \frac{1}{1 + \mathbf{a}^\top \mathbf{W}\mathbf{a}/\mathbf{a}^\top \mathbf{B}\mathbf{a}}.$$

This vector is unique up to a constant.

Equivalent Formulations of the Problem

The discriminant analysis problem, as formulated by Fisher, is then equivalent to finding a vector $\mathbf{a} \in \mathbb{R}^p$ such that

- ✓ $\mathbf{a}^\top \mathbf{B}\mathbf{a}/\mathbf{a}^\top \mathbf{S}\mathbf{a}$ is maximized under the constraint that $\mathbf{a}^\top \mathbf{a} = 1$

or

- ✓ $\mathbf{a}^\top \mathbf{B}\mathbf{a}$ is maximized under the constraint that $\mathbf{a}^\top \mathbf{S}\mathbf{a} = 1$.

Upon setting $\mathbf{c} = \mathbf{S}^{1/2}\mathbf{a}$, the problem is also equivalent to

$$\boxed{\text{maximize } \mathbf{c}^\top \mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}\mathbf{c} \text{ under the constraint that } \mathbf{c}^\top \mathbf{c} = 1.}$$

Solution of the Optimization Problem

The third formulation of the problem is to maximize

$$\mathbf{c}^\top(\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2})\mathbf{c},$$

under the constraint that $\mathbf{c}^\top\mathbf{c} = 1$.

This problem was already solved in PCA! The solution is to take

$$\mathbf{a} = \mathbf{S}^{-1/2}\mathbf{c},$$

where $\mathbf{c}$ is a normed eigenvector corresponding

$$\lambda = \text{largest eigenvalue of } \mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}.$$

Equivalent Solution

Equivalently, from the second formulation and the chapter on correspondence analysis, one can take

$$\mathbf{a} = \text{normed eigenvector}$$

associated with

$$\lambda = \text{largest eigenvalue of } \mathbf{S}^{-1}\mathbf{B}.$$

Note that if

$$\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}\mathbf{c} = \lambda\mathbf{c} \quad \text{and} \quad \mathbf{a} = \mathbf{S}^{-1/2}\mathbf{c},$$

then

$$\mathbf{S}^{-1/2}\mathbf{B}\mathbf{a} = \lambda\mathbf{S}^{1/2}\mathbf{a} \quad \Rightarrow \quad \mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}.$$

Therefore, the eigenvalues of $\mathbf{S}^{-1}\mathbf{B}$ and $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$ are the same.