SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

J. MacQUEEN
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

The main purpose of this paper is to describe a process for partitioning an N-dimensional population into k sets on the basis of a sample. The process, which is called 'k-means,' appears to give partitions which are reasonably efficient in the sense of within-class variance. That is, if p is the probability mass function for the population, $S = \{S_1, S_2, \cdots, S_k\}$ is a partition of E_N, and u_i, $i = 1, 2, \cdots, k$, is the conditional mean of p over the set S_i, then $w^2(S) = \sum_{i=1}^{k} \int_{S_i} |z - u_i|^2 \, dp(z)$ tends to be low for the partitions S generated by the method. We say 'tends to be low,' primarily because of intuitive considerations, corroborated to some extent by mathematical analysis and practical computational experience. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer. Possible applications include methods for similarity grouping, nonlinear prediction, approximating multivariate distributions, and nonparametric tests for independence among several variables.

In addition to suggesting practical classification methods, the study of k-means has proved to be theoretically interesting. The k-means concept represents a generalization of the ordinary sample mean, and one is naturally led to study the pertinent asymptotic behavior, the object being to establish some sort of law of large numbers for the k-means. This problem is sufficiently interesting, in fact, for us to devote a good portion of this paper to it. The k-means are defined in section 2.1, and the main results which have been obtained on the asymptotic behavior are given there. The rest of section 2 is devoted to the proofs of these results. Section 3 describes several specific possible applications, and reports some preliminary results from computer experiments conducted to explore the possibilities inherent in the k-means idea. The extension to general metric spaces is indicated briefly in section 4.

The original point of departure for the work described here was a series of problems in optimal classification (MacQueen [9]) which represented special

This work was supported by the Western Management Science Institute under a grant from the Ford Foundation, and by the Office of Naval Research under Contract No. 233(75), Task No. 047-041.