PARTITIONING A DISTRIBUTION IN ONE OR TWO DIMENSIONS

Shyam Johari and Stanley L. Sclove, University of Illinois at Chicago Circle

ABSTRACT

Suppose it is desired to partition a distribution into k groups (k given) using squared error or absolute error as the measure of information retained. An algorithm to obtain the optimal boundaries (or class probabilities) is given. Optimal class probabilities for the case of squared error, were obtained for k = 2 to 15 for beta (for various values of the parameters), chi-square (12 d.f.), exponential, normal, and uniform distributions. Results obtained are compared and analyzed in the light of existing papers.

The problem of partitioning a bivariate normal distribution into k groups using squared error as the measure of information retained is also considered. One way of formulating this mathematically is given and the results so obtained are discussed.

Problems of partitioning subject to loss functions incorporating both squared error and a cost associated with the number of classes are considered. Some problems for further investigation are outlined.

In these Proceedings we only summarize in words the results obtained. Mathematical statements of the results, proofs, and tables are given in Johari(1975).

1. INTRODUCTION

Let X be a numerical random variable measuring some property which is conveniently represented by ordered categories. E.g., X may be some numerical variable related to health, and it may be convenient to classify people as in poor, fair, or good health.

By partitioning the distribution of the random variable X we mean assigning each of its possible values to one of k classes. The problem is to define the classes, or, equivalently, to determine the class probabilities or boundaries. The problem of partitioning is known also as the grouping problem.

The problem of assigning letter grades in courses may be considered in these terms. An instructor is confronted with the task of giving one of the five letter grades A, B, C, D, or F to each of the students in his class on the basis of their measured performance, X. Sometimes the distribution (10%,20%,40%,20%,10%) is used. If one used a partition of the normal distribution given by boundaries one standard deviation apart, that is, at -1.5, -0.5, +0.5, +1.5, then the distribution over A,B,C,D,F would be (6.68%, 24.17%, 38.30%, 24.17%, 6.68%). Rounding to the nearest 10% produces the distribution (10%, 20%, 40%, 20%, 10%).

During World War II the Army Air Force Aviation Psychology Program used a stanine scale for scoring psychological tests. [See, e.g., Ferguson(1971), p.386.] The possible range of scores on a test was partitioned into nine intervals, labelled 1, 2, ..., 9. These scores are called stanines (short for "standard nine"). Stanine scores correspond to intervals of equal width along the X-axis. The width used is one-half a standard deviation unit. A stanine of 5 covers the interval from -.25 to +.25 in standard deviation units; a stanine of 6, the interval +.25 to +.75; etc. A stanine of 9 includes not only the interval 1.75 to 2.25 but also all cases above 2.25; similarly, a stanine of 1 includes all cases below -2.25. The corresponding distribution is (4%, 7%, 12%, 17%, 20%, 17%, 12%, 7%, 4%). The question arises as to how good this grouping may be. The question also arises as to how much additional information is conveyed by say, 9 groups instead of 8.

Given some reasonable measure of goodness of a grouping, what is the best grouping into k classes for a normal distribution? Cox (1957) considered this question, using a measure of goodness corresponding to squared error loss and obtaining explicit numerical results for k = 2 to 6. Connor (1972) obtained results, again for k = 2 to 6, for the exponential distribution. He focused on the problem of maximizing the asymptotic efficiency of a test for correlation of a binary variable with the continuous variable X. He notes various contexts in which the mathematical problem equivalent to that of determining the optimal classes appears.

For some purposes, say for ease of calculation or graphical representation, it may be best to use groups of equal width, as in the case of stanines. Given k, what is the best equal-width grouping for a normal distribution? Stefansky and Kaiser (1973) considered this question. Using a measure of goodness corresponding to squared error loss, they give optimal equal-width groupings for a normal distribution for k = 2 to 15.

Having summarized the previous work in this area, we shall now present our results. In these Proceedings we only summarize in words the results obtained. Mathematical statements of the results, their proofs, and tables are given in Johari(1975).

2. PARTITIONING A UNIVARIATE DISTRIBUTION

2.1. Case of Squared Error

Consider a continuous probability density function with support S, which may be finite or infinite, and finite variance, σ^2. Suppose it is desired to parti-ion S into k intervals. With th i-th interval associate a value v(i) and imagine that each individual put into the i-th interval is given the value v(i). That is, we have a random variable v(X), a function of the random variable X, defined by v(x) = v(i) when X is in the i-th interval. We measure the loss due to assigning an individual with the value x to the i-th interval by $[x - v(i)]^2/\sigma^2$.

The expected loss L due to partitioning is the expected value, $L = E[X - v(X)]^2/\sigma^2$. To minimize L, proceed in two stages. First, for any fixed partition, choose v(1), ..., v(k) to minimize L; then choose the partition. To achieve the first part of the minimization, choose v(i) to be the conditional mean of the i-th interval, since the second moment about the mean is less than that about any other point. From now on v(i) will be used to denote the conditional mean of the i-th interval. Also, v will denote E(X).

If p(i) denotes the probability that an observation falls in the i-th interval, then $\sigma^2 L = \Sigma p(i) E[(X - v(i))^2 | X \text{ in i-th interval}]$, where the summation is over 1,2,...,k. Note that $L = 1 - M$, where $M = \Sigma\, p(i)[v(i) - v]^2/\sigma^2$ is a normalized between-groups sum of squares. We wish to mazimize M. Observe that if k = 1, then L = 1, representing complete loss of information about differences among individuals.

2.1.1. Characterization of the Boundaries

Theorem 1. Each of the optimal boundaries is halfway between the conditional means of the adjacent intervals of the partition.

Theorem 2. If the distribution is symmetric , then the optimal boundaries are symmetric about the mean.

2.1.2. An Algorithm for Constructing the Boundaries

Step 1: Pick any k-1 boundaries to begin.

Step 2: Find the conditional means of the corresponding intervals.

Step 3: Compute new boundaries, the points halfway between the conditional means.

Stop when sufficient accuracy has been attained.

2.1.3. Numerical Results

Optimal partitions were obtained for the following distributions: (i) Normal, (ii) Exponential, (iii)Uniform, (iv) Chi-square(12 d.f.) and (v) Beta with parameters m=10, n=2; m=2,n=2; m=.5, n=1; m=5, n=5; and m=3, n=2, using the algorithm on an IBM 370/155. The results are tabulated in Johari (1975).

It should be observed that once the optimal boundaries are known for say the standard normal distribution N(0,1), the boundaries for $N(\mu,\sigma^2)$ can be obtained by replacing a boundary B by $\mu + \sigma B$. The percentages of individuals in each group remain the same. The transformation for the exponential distribution with mean μ=1 to μ=m is given by replacing B by mB. This works for any location-scale parameter family. This is not the case for chi-square and beta distributions.

Throughout the analysis we assumed the mean and the standard deviation of the distribution to be known. If instead we are given a random sample from a distribution in a specified family, the parameters being unknown, we estimate them in the usual way and apply the results to the estimated distribution.

2.1.4. Discussion and Applications

a. Comparison of results for various distributions. From the numerical results we observed that if we deal with distributions with less important tails, it is best to allow the frequency in the tail groups to rise, whereas in a distribution with one long tail, it is best to have a tail group or groups with lower frequencies. This agrees wiht what Cox (1957) conjectured. If the distribution is symmetric, then as proved in Theorem 2, it is best to allow the frequency of the first group to be equal to the frequency of the last group, the frequency of the second group to be equal to that of the second to last group, etc. We also observed that the normal distribution appears to require at least as may classes to retain a specified amount of information as does any other distribution. This is analogous to the fact that the normal is the maximum entropy distribution, among those with given variance. [See,e.g., Ash(1965), p. 240.]

b. Number of groups. The numerical results show that the effect of increasing the number of classes is small beyond three or four classes. Therefore, these results suggest that it is advisable to use more than two classes, but that using more than four or five yields only marginally more information. Moreover, the classes are quite robust with respect to moderate departure from their optimal probabilities. E.g., equal frequencies partitions retain almost as much information as optimal partitions.

For the problem of assigning letter grades in courses, the case k = 2 relates to a pass-fail system; k = 3 is what one often encounters in grading graduate students; k = 4 (A,B,C,D) and k = 5 (A,B,C,D,F) occur typically in grading undergraduates. The cases of k = 6 to 15 occur where pluses and minuses are attached to the letter grades.

c. Case of five groups. The case of five groups is of special interest because of the widespread use of the letter grades (A,B,C,D,F). Some schools use the distribution (10%, 20%, 40%, 20%, 10%) over these grades. This distribution is also recommended by at least some accreditation organizations, for example the Engineers' Council for Professional Development. This distribution is nice because all the percentages are multiples of 10. If there are n students, then m(A) = .10n is the number receiving A, m(B) = 2m(A) is the number receiving B; m(C) = 2m(B) is the number receiving C; m(D) = m(B); m(F) = m(A).

As was pointed above, use of boundaries one standard deviation apart produces the distribution (6.68%, 24.17%, 38.30%, 24.17%, 6.68%). Rounding the percentages in this distribution to the nearest 10% produces the distribution (10%,20%,40%,20%,10%). Rounding the percentages to the nearest 5% produces the distribution (5%,25%, 40%,25%,5%). This differs from (10%, 25%, 30%, 25%, 10%), the distribution obtained by rounding the percentages of the optimal distribution to the nearest 5%, which we recommend instead of (10%,20%,40%,20%,10%).

d. Near optimality of stanines

For the stanine scale, the Army Air Force used intervals of width .5, which correspond to the distribution (4%,7%,12%,17%,20%,17%,12%,7%,4%). Stefansky and Kaiser(1973) consider the problem of partitioning the normal distribution into k = 2 to 15 groups, minimizing the squared error loss, subject to the condition of equally spaced boundaries. They state the optimal width for k = 9 to be .5338, which is not in agreement wiht our results. A width of .5338 retains only 97.06% of the information, which is indeed worse than the stanines' width of .5, which retains 97.07%. (Stefansky and Kaiser have 96.93% instead of 97.06%.) According to our calculations the width should be taken to be .51 (retaining 97.08% of the information), very close indeed to what the Army Air Force used. The overall best distribution for k = 9, accorindg to our calculations, is (3%,9%,13%,16%,18%, 16%,13%,9%,3%). The information retained in this case is 97.21%.

3. CASE OF ABSOLUTE ERROR

2.2 Case of Absolute Error

Analogous results, with medians replacing means, hold when the loss function is absolute error.

3. PARTITIONING A BIVARIATE NORMAL DISTRIBUTION

Let $\underline{X}$ denote a bivariate continuous random variable, i.e., $\underline{X}' = (X_1, X_2)$ with support S (which is a subset of the plane. Suppose it is desired to partition S into k groups. With the i-th group, S(i), we associate a vector $\underline{v}(i)$ and imagine that each object put into the i-th group is given the value $\underline{v}(i)$. I.e., we have a random variable $\underline{v}(\underline{X})$, a function of the random vector $\underline{X}$, defined by $\underline{v}(\underline{x}) = \underline{v}(i)$, when $\underline{x}$ is in S(i). The loss due to grouping an object with value $\underline{x}$ into the i-th group is measured by $[\underline{x} - \underline{v}(i)]'\Sigma^{-1}[\underline{x} - \underline{v}(i)]$, vector squared error in the metric of Σ, the covariance matrix of $\underline{X}$. The loss function L is taken to be half the expectation of this squared error. The factor of two is used so that L =1 when k=1. Again, the minimization can be done in two stages, the $\underline{v}(i)$ being taken to be the conditional mean vectors.

Now assume $\underline{X}$ is bivariate normal. Furthermore, let us assume that the set S(i) is the sector between the angles $\theta(i)$ and $\theta(i-1)$ (in polar co-ordinates), with the last angle being 2π and the first being 0. Writing L in terms of the angles, setting the partials of L with respect to the angles equal to zero, and solving shows that the unique solution is given by $\theta(i) = (2\pi/k)i + \theta(0)$ for i = 1,2, ...,k-1. This proves the theorem: Let $\underline{X}$ be distributed according to $N(\underline{0},\underline{I})$, the bivariate normal distribution with mean vector $\underline{0}$ and covariance matrix I. Suppose that is is desired to partition this distribution into k groups, where each group

is wedge-shaped, the wedges having their vertices at (0,0), the i-th group being given by $(\theta(i-1),\theta(i))$ in polar co-ordinates, using L as the measure of information retained. Then the unique optimal partition is given by the equi-angular partition.

Results for the problem of partitioning when the covariance matrix is not the identity and the loss function is squared error in the metric of the covariance matrix can be obtained by transforming to the case in which the covariance matrix is the identity. The results show that smaller angles are needed where the density is high.

4. SOME EXTENSIONS

4.1. Loss Functions Incorporating the Number of Groups

Given k, the number of groups, we can find the optimal partition of a given distribution. We now impose additional conditions to force the existence of a single best k, that is, we add a function of k to the loss. Consider the loss function,
Loss = Squared error + Ck,
where C is some positive constant, the cost per group. If C is small, a large value of k is optimal; if C is large, a small value of k is optimal. The optimal k is a non-increasing step function of C. Values are given in Johari(1975).

4.2. Loss Function Incorporating Cost of Inequity of Partition

Suppose it is physically convenient if the groups have the same probability. (One can think of some such problem as assigning pupils to k teachers, using some dimension of ability, where it would be convenient if each teacher had the same number of pupils.) Then it makes sense to add to the original loss function a term measuring the inequity of the partition.

We used a function analogous to the "chi-square" statistic for measuring the departure of observed relative frequencies from expected relative frequencies, in this case, a uniform distribution. In this case one could find the optimal k by computer. If the weight given the chi-square portion of the loss is small, the solution will be close to that obtained without consideration of the cost of inequity. If this weight is large, a distribution close to the uniform will be optimal.

REFERENCES

Ash, Robert(1965). Information Theory. Interscience Publishers, New York.

Connor, Robert J.(1972). "Grouping for Testing Trends in Categorical Data," Journal of the American Statistical Association, 67, pages 601-4.

Cox, D.R.(1957). "Note on Grouping," Journal of the American Statistical Association, 52, pages 543-7.

Ferguson, George A.(1971). Statistical Analysis in Psychology and Education, 3rd edition, McGraw-Hill, New York.

Johari, Shyam(1975). "Partitioning a Distribution in One or Two Dimensions," Ph.D. thesis, Department of Mathematics, University of Illinois at Chicago Circle.

Stefansky, W., and Kaiser, Henry F. (1973). "Note on Discrete Approximations," Journal of the American Statistical Association, 68, pages 232-4.