A MATHEMATICAL THEORY OF PATTERN RECOGNITION

BY ARTHUR ALBERT


1. Summary. Let $X_0$ and $X_1$ be (unknown) disjoint subsets of a Hilbert space $H$, such that the convex hulls of $X_0$ and $X_1$ are a positive distance apart. Suppose that samples are drawn independently and at random from $X_0 \cup X_1$. After the $n$th sample, $Z_n$, it is required to guess whether $Z_n$ came from $X_0$ or $X_1$. After each guess, we are told whether we were right or wrong.

In this paper, a decision procedure is exhibited, having the property that the probability of making an error on the $n$th trial converges to zero with increasing $n$. Furthermore, the guessing rule used on the $n + 1$st trial depends on the past data only through the rule used on the $n$th trial, the value of $Z_n$, and whether or not the guess about $Z_n$ was correct.

The application to pattern recognition problems of a dichotomous sort is immediate when we identify $X_0$ and $X_1$ with two classes of patterns which are observed in temporal succession. The rules for membership in $X_0$ and $X_1$ are not known, but we (or a machine) are/is told to which class each pattern belongs, after making a guess about that pattern. As the “training period” increases, errors are made with ever decreasing frequency.

2. Mathematical formulation. Representative samples are to be drawn independently and at random from each of two classes (call the classes “class zero” and “class one”). Associated with each phenomenon to be classified is a measurable attribute. The set of attributes corresponding to members of “class zero” will be denoted by $X_0$, while those corresponding to members of “class one” will be denoted by $X_1$. Suppose for the moment that $X_0$ and $X_1$ are non-overlapping sets and that there is a real valued function defined over $X_0 \cup X_1$ having the property that, for some constant, $c$,

$$\inf_{x \in X_0} f(x) > c > \sup_{x \in X_1} f(x).$$

If we did not know the rule for membership in $X_0$ or $X_1$, but we did know $f$ and $c$, we could still carry out the decision procedure viz: After each observation, apply $f$ to the attribute $x$. If $f(x) > c$, decide $x \in X_0$ (and thus the phenomenon belongs to “class zero”). Otherwise, decide $x \in X_1$.

Suppose $f$ and $c$ are not known: If we are told after each decision whether the attribute really belonged in $X_0$ or $X_1$, it seems reasonable to hope that, under certain circumstances, this information can be utilized in some way to make estimates of $f$ and $c$. If we are clever enough to construct estimates of $f$ and $c$ that converge, in an appropriate sense, to the true values as more data

Received January 22, 1962.

1 This work was initiated during the summer of 1961, while the author was employed at Laboratory for Electronics, Brighton, Massachusetts.

284