From Proceedings of FIRST INTERNATIONAL JOINT CONFERENCE ON PATTERN RECOGNITION, (WASH. D.C., OCT, 1973)

GENERALIZATION ON PATTERNS USING KOLMOGOROV COMPLEXITY

Thomas M. Cover

Stanford University
Stanford, California

0. Summary

Given a domain D of patterns $D = \{x_1, x_2, \ldots x_n\}$ and an unknown classification function $f: D \to \{0,1\}$ assigning the patterns to two classes, we ask for an intelligent way to learn f as the correctly classified elements in D are presented one by one. We ask this question in a gambling context in which a gambler, starting with one unit, sequentially bets a portion of his current capital on the classification of the new pattern. We find the optimal gambling system when f is known a priori to belong to some family F. We also exhibit a universal optimal learning scheme achieving $\exp_2(n - K(f|D) - \log(n+1))$ units for each f, where $K(f|D)$ is the length of the shortest binary computer program which calculates f on its domain D. In particular it can be shown that a gambler can double his money approximately $n(1 - H(d/n))$ times, where $H(p) = -p \log p - (1-p)\log(1-p)$, if f turns out to be a linear threshold function on n patterns in d-space.

1. Kolmogorov Complexity

Let N denote the natural numbers $\{0,1,2,\ldots\}$. Let $x \in \{0,1\}^\infty$ denote an infinite binary sequence $x = (x_1, x_2, \ldots)$ and let $x(n) = (x_1, x_2, \ldots, x_n)$ denote the first n terms. Let $\{0,1\}^*$ denote all binary sequences of finite length. Let A be a recursive function $A: \{0,1\}^* \times N \to \{0,1\}^*$. Let $\ell(x)$ denote the length of the sequence x. Then $K_A(x(n)|n) = \min_{A(p,n)=x(n)} \ell(p)$ is defined to be the complexity of x(n) with respect to the algorithm A, given the length n of the sequence x(n). If A is a universal recursive function, then K_A, or simply K, is called the Kolmogorov complexity. We know that

i) $K(x(n)|n) \leq K_B(x(n)|n) + C_B$ for all $n \in N$, $\forall x$

ii) $|\{x \in \{0,1\}^* : K(x) = k\}| \leq 2^k$, $\forall k \in N$.

Now we define a complexity measure for functions $f: D \to \{0,1\}$, where the domain D is some finite set. Let A be a universal recursive function.

Def. $$K_A(f|D) = \min_{\substack{A(p,x)=f(x) \\ \forall x \in D}} \ell(p)$$

Thus the complexity of f given the domain D is the minimum length program p such that a Turing machine A, or equivalently a mechanical algorithm A, can compute f(x) in finite time, for each $x \in D$.

Example 1. Let $D = \{0,1\}^d$.

$$\text{Let} \quad f(x) = \begin{cases} 1, & \sum x_i = \text{odd} \\ 0, & \sum x_i = \text{even} \end{cases}$$

Then $K(f|D) = c$, where c is some small constant independent of d. The parity function above is easy to describe and thus has essentially zero complexity.

Example 2. Let $D = \{(p,q)\}$ be the set of all rational numbers p/q. Let

$$f(p,q) = \begin{cases} 1, & |\pi - p/q| > \frac{1}{q^{51}} \\ 0, & \text{otherwise} \end{cases}$$

Then f has complexity equal to some small constant. In fact, by a theorem due to Mahler, $f(p,q) \equiv 1$, $\forall (p,q) \in D$. Thus f has trivial complexity. Some f's of nontrivial complexity will be given in the applications of the main theorem.

2. The Main Theorem

Let F denote a set of (classification) functions $f: D \to \{0,1\}$. For example F might be the set of all linear threshold functions. Let $|F|$ denote the number of elements in F.

The interpretation will be that D is the set of patterns, and f(x) is the classification of the pattern x in D.

Consider the following gambling situation. The elements of D are presented in any order. A gambler starts with one dollar. The first pattern $x_1 \in D$ is exhibited. The gambler then announces amounts b_1 and b_0 that he bets on the true class being $f(x_1) = 1$ and $f(x_1) = 0$, respectively. Without loss of generality we can set $b_1 + b_0 = 1$. The true value $f(x_1)$ is then announced, and the gambler loses the incorrect bet and is paid even money on the correct bet. Thus his new capital is

$$S_1 = \begin{cases} 2b_1, & f(x_1) = 1 \\ 2b_0, & f(x_1) = 0 \end{cases}.$$

Now a new pattern element $x_2 \in D$ is exhibited. Again the gambler announces proportions b_1 and b_0 of his current capital that he bets on

$f(x_2) = 1$ and $f(x_2) = 0$ respectively. Without loss of generality let $b_0 + b_1 = 1$. Thus the bet sizes are b_1S_1 and b_0S_1. Then $f(x)$ is announced and the gambler's new capital is

$$S_2 = \begin{cases} 2b_1S_1, & f(x_2) = 1 \\ 2b_0S_2, & f(x_2) = 0 \end{cases} .$$

Continuing in this fashion, we define

$$b_1^{(k)}\left(x_k \mid (x_1,f(x_1)),\ldots,(x_{k-1},f(x_{k-1}))\right),\ x_k \epsilon D$$

and $b_0^{(k)} = 1-b_1^{(k)}$, $b_0^{(k)} \geq 0$, $b_1^{(k)} \geq 0$ as a gambling scheme that depends only on the previously observed properly classified (training) set.

The accrued capital after all patterns $x_1,x_2,\ldots,x,n = |D|$, have been observed is

$$S_k = \begin{cases} 2b_1^{(k)}S_{k-1}, & f(x_k) = 1 \\ 2b_0^{(k)}S_{k-1}, & f(x_k) = 0 \end{cases}$$

for $k = 1,2,\ldots,n$ and $S_0 = 1$. Let

$$b = \left(\left(b_0^{(1)},b_1^{(1)}\right),\ \left(b_0^{(2)},b_1^{(2)}\right),\ldots,\left(b_0^{(n)},b_1^{(n)}\right)\right)$$

denote a sequence of gambling functions.

Theorem 1. For any $F \subseteq D^{\{0,1\}}$, there exists a gambling scheme b^* achieving $S_n(f) = S^* = 2^{n-\log|F|}$ units, for all f in F and for all orders of presentation of the elements $x \epsilon D$. Moreover, there exists no b that dominates b^* $\forall f$; thus b^* is minimax. This gambling scheme is given by the expression

$$b_1^{(k)*}(x) = \frac{|\{g \epsilon F : g(x_i) = f(x_i),\ i = 1,2,\ldots,k-1,\text{ and } g(x) = 1\}|}{|\{g \epsilon F; g(x_i) = f(x_i),\ i = 1,2,\ldots,k-1\}|}$$

Remark. This gambling scheme simply asserts at time k, "Bet all of the current capital on the hypotheses $f(x_k) = 1$ and $f(x_k) = 0$ in proportion to the number of functions g in F that agree on the training set and assign the new pattern x_k to classes $g(x_k) = 1$ and $g(x_k) = 0$ respectively."

The proof will not be given here but can be found in [1].

Applications and Examples.

1. Let F be all 2^n functions $f:D \to \{0,1\}$, where $n = |D|$. Then $\log|F| = n$, and $S^* = 1$. No money can be gained. The expanding training set gives no information about future pattern classifications. This is the worst case.

2. Let D denote a set of n vectors in Euclidean d-space R^d. Let us also assume that $\{x_1,x_2,\ldots,x_n\} = D$ is in general position in the sense that every d-element subset of D is linearly independent. Let F be the set of all linear threshold functions on D; i.e., $f \epsilon F$ implies there exists $w \epsilon R^d$, $T \epsilon R$, such that

$$f(x) = \text{sgn}(w^t x - T),\ \forall x \epsilon D,$$

where

$$\text{sgn}(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 . \end{cases}$$

Then from Cover [1965], we have $|F| = 2\sum_{k=0}^{d}\binom{n-1}{k}$, $\forall d,n$. Using bounds derived from Stirling's approximation, it can be shown that

$$\log\left(2\sum_{k=0}^{d}\binom{n-1}{k}\right) \approx nH\left(\frac{d}{n}\right), \quad \text{for } n \geq 2d,$$

where $H(p) = -p \log p - (1-p)\log(1-p)$ is the Shannon entropy function.

Thus we conclude, for $n \geq 2d$, that an amount $S_n = 2^{n(1 - H(d/n))}$ can be won if in fact the n patterns are linearly separable in R^d. Note also that H(d/n) is the Kolmogorov complexity of most of the linear threshold function $f \epsilon F$. Finally, we observe that S_n is not much greater than 1 until $n \geq 2d$, at which point the behavior of S_n is exponential. This is yet more evidence that n = 2d is a natural definition of the capacity of a linear threshold pattern recognition device with d variable weights.

3. Let F be the set of all functions $f:D \to \{0,1\}$ that can be represented by rth degree polynomial discriminant functions:

$$f(x) = \text{sgn}\left[\sum_{i_1,i_2,\ldots,i_r} w_{i_1i_2\ldots i_r} x_{i_1}x_{i_2}\ldots x_{i_r} - T\right]$$

If the elements of D are in general position with respect to rth degree polynomials, we see [2] that there are precisely

$2\sum_{k=0}^{d'-1}\binom{n-1}{k}$ elements in F where d' is the number of coefficients in an arbitrary r^{th} degree polynomial in d variables. For example, for r = 2, d = 2, we have $f(x) = \left[a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2 + a_1x_1 + a_2x_2 + a_0\right]$, and d' = 6. The point is that d' is the number of degrees of freedom of the manifold $\{x:f(x) = 0\}$. Again by the theorem we have $S_n \geq 2^{n(1 - H(d'/n))}$, where now d' is the number of degrees of freedom of the family of separating surfaces F.

4. Suppose it is not known what degree polynomial is needed to classify D correctly. Since the degree r need take on only (n+1) values before the degree is sufficient to make an arbitrary assignment f, we merely invest an initial amount 1/(n+1) in the betting system for each degree r = 0,1,...,n. Then the theorem becomes $S(f) \geq 2^{n(1 - H(d(f)/n)) - \log(n+1)}$, for all $f:D \to \{0,1\}$ where d(f) is the number of degrees of freedom of an r^{th} degree polynomial, where r is the minimal degree necessary to yield f.

Final Theorem. These results are special cases of the following theorem:

Theorem: $\exists$ a betting scheme b* such that the total accumulated capital satisfies $S(f) \geq 2^{n - K(f|D) - \log(n+1)}$.

Comment. If f is a linear threshold function, then

$$K(f|D) \leq \log 2\left[\sum_{k=0}^{d-1}\binom{n-1}{k}\right] + C .$$

Simply write a program saying "f is a linear threshold function, and is the i^{th} on the list of functions arranged in lexicographic order." Thus i requires $\log 2\left[\sum_{k=0}^{d-1}\binom{n-1}{k}\right]$ bits and c is the length of the rest of the program specified above.

Similarly, the polynomial threshold functions can be seen to be special cases of the final theorem.

References

1. Cover, Thomas M., "Universal Gambling Schemes and Kolmogorov Complexity," in preparation.

2. Cover, Thomas M., "Geometrical and Statistical Properties of Linear Threshold Functions with Applications in Pattern Recognition," IEEE Trans. Elec. Comp., 1965.