特徴空間の次元数と学習パターン数

2010. 5. 11　石井

合計 n 個のパターンが d 次元特徴空間上に分布しており、各パターンはクラス ω_1, ω_2 のいずれかに属するものとする（$c = 2$)。各パターンは独立に ω_1 または ω_2 のいずれかに属するので、n 個のパターンの分布の場合の数は 2^n である。このうち、線形識別可能な場合の数を $L(n, d)$ とすると、

$$L(n, d) = (d \text{ 次元特徴空間上の } n \text{ 個のパターンを二分できる場合の数}) \times 2 \tag{1}$$

が成り立つ。例えば、$n = 4, d = 2$ とすると、

$$\begin{aligned} L(n, d) &= L(4, 2) && (2)\\ &= 7 \times 2 && (3)\\ &= 14 && (4) \end{aligned}$$

である。この $L(n, d)$ を求めるには漸化式

$$L(n, d) = L(n-1, d) + L(n-1, d-1) \tag{5}$$

を用いる（証明は省略)。ここで初期条件

$$\begin{aligned} L(1, d) &= 2 && (6)\\ L(n, 1) &= 2n && (7) \end{aligned}$$

が成り立つことは明らかであるので、数学的帰納法により、

$$L(n, d) = \begin{cases} 2 \cdot \displaystyle\sum_{j=0}^{d} {}_{n-1}C_j & (n \geq d+1) \\ 2^n & (n \leq d) \end{cases} \tag{8}$$

が得られる。例えば、$n = 4, d = 2$ とすると、

$$\begin{aligned} L(n, d) &= L(4, 2) && (9)\\ &= 2 \cdot \sum_{j=0}^{2} {}_3C_j && (10)\\ &= 2 \cdot ({}_3C_0 + {}_3C_1 + {}_3C_2) && (11)\\ &= 2 \cdot (1 + 3 + 3) && (12)\\ &= 14 && (13) \end{aligned}$$

となり、式 (4) と一致することが確かめられる。

以上の結果より、d 次元空間上の n 個のパターンが、2 つのクラス ω_1, ω_2 のいずれかに属するとき、これらを線形分離できる確率 $P(n,d)$ は、

$$P(n,d) = L(n,d)/2^n \tag{14}$$

$$= \begin{cases} 2^{1-n} \cdot \sum_{j=0}^{d} {}_{n-1}C_j & (n \geq d+1) \\ 1 & (n \leq d) \end{cases} \tag{15}$$

となる。ここで、

$$\lambda \stackrel{\text{def}}{=} n/(d+1) \tag{16}$$

として、λ と $P(n,d)$ の関係をプロットしてみると、$\lambda = 2$ のとき $n = 2(d+1)$ であるから、式 (15) より、

$$P(n,d) = 2^{-2d-1} \cdot \sum_{j=0}^{d} {}_{2d+1}C_j \tag{17}$$

$$= 2^{-2d-1} \cdot 2^{2d} \tag{18}$$

$$= 1/2 \tag{19}$$

となる。すなわち、パターン数が次元数の約 2 倍のとき、線形分離できる確率は 1/2 であることがわかる。また、次元数 d が大きい場合は、$\lambda = 2$ 近辺で閾値効果がある（教科書 65p 式（4.67）参照)。

なお、式 (17) から式 (18) の導出には、

$$\sum_{j=0}^{d} {}_{2d+1}C_j = \frac{1}{2} \sum_{j=0}^{2d+1} {}_{2d+1}C_j \tag{20}$$

および、二項定理

$$(x+y)^n = \sum_{k=0}^{n} {}_nC_k \, x^{n-k} y^k \tag{21}$$

において $x = y = 1$ とおくことにより得られる

$$\sum_{k=0}^{n} {}_nC_k = 2^n \tag{22}$$

を用いた。