Информатика

Лекция 3
Дискретные источники

Дискретные источники без памяти

- Дискретный источник без памяти (ДИБП) полностью задается таблицей распределения вероятностей:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_N \\ p_1 & p_2 & \dots & p_N \end{pmatrix}$$

- Информационные характеристики ДИБП:

Энтропия $H(X) = -\sum_{i=1}^{N} p_i \log p_i$; максимальная энтропия $H_{\max} = \log N$

Избыточность $D = H_{\max} - H(X)$;

Относительная избыточность $D_0 = \frac{D}{H_{\max}} = \frac{H_{\max} - H(X)}{H_{\max}}$;

Средняя длительность выдачи одного символа $\tau_{cp} = \sum_{i=1}^{N} p_i \, \tau_i$;

Производительность источника $\bar{I}(X) = \frac{H(X)}{\tau_{cp}}$.

Дискретные источники с памятью

Дискретный источник с памятью r и алфавитом $x_1, x_2, \ldots x_N$ задается множеством состояний $\{S_1, S_2, \ldots S_M\}$, $M = N^r$; вероятностями $p^{(j)}(i)$ появления символа x_i в состоянии S_j; графом перехода из состояния $S[n] = (x[n-r], x[n-r+1] \ldots x[n-1])$ в $S[n+1] = (x[n-r+1], x[n-r+2] \ldots x[n])$ после появления очередного символа $x[n]$; начальным распределением состояний $\mathbf{p}_0 = (p_0(1), p_0(2) \ldots p_0(M))$.

Будем предполагать стационарность (независимость от начала отсчета времени) и эргодичность (однотипность всех возможных последовательностей символов). Любой стационарный источник можно представить как несколько эргодических источников, различающихся режимами работы.

Совместные энтропии возрастающих последовательностей символов: $H(X_1X_2)$, $H(X_1X_2X_3)$, дают в среднем на один символ энтропию

$$H_L(X) = \frac{1}{L} H(X_1X_2 \ldots X_L).$$

Условная энтропия L-го символа $H(X_L \mid X_1X_2 \ldots X_{L-1})$.

Теорема. $\lim\limits_{L\to\infty} H(X_L \mid X_1X_2 \ldots X_{L-1}) = \lim\limits_{L\to\infty} H_L(X) = H_\infty(X)$.

Цепи Маркова

Математической моделью ДИСП является цепь Маркова. Это последовательность состояний $S[1], S[2], \ldots, S[n], \ldots$, каждое из которых принадлежит множеству $\{S_1, S_2, \ldots S_M\}$, и заданные вероятности перехода $\pi_{n_2 n_1}(j_2 \mid j_1) = P(S[n_2] = S_{j_2} \mid S[n_1] = S_{j_1})$.

Они должны удовлетворять следующим свойствам:

1. $\sum_{j_2=1}^{M} \pi_{n_2 n_1}(j_2 \mid j_1) = 1$ (полная группа);
2. $\pi_{n_3 n_1}(j_3 \mid j_1) = \sum_{j_2=1}^{M} \pi_{n_3 n_2}(j_3 \mid j_2) \cdot \pi_{n_2 n_1}(j_2 \mid j_1)$ (равенство Колмогорова-Чэпмена).

Для стационарного ДИСП (независимого от начала отсчета времени) достаточно задать

$\pi_{n_1+l,n_1}(j_2 \mid j_1) = P(S[n_1+l] = S_{j_2} \mid S[n_1] = S_{j_1})$, то есть матрицу

$$\Pi = \begin{pmatrix} \pi_{n+1,n}(1|1) & \pi_{n+1,n}(2|1) & \cdots & \pi_{n+1,n}(M|1) \\ \pi_{n+1,n}(1|2) & \pi_{n+1,n}(2|2) & \cdots & \pi_{n+1,n}(M|2) \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{n+1,n}(1|M) & \pi_{n+1,n}(2|M) & \cdots & \pi_{n+1,n}(M|M) \end{pmatrix},$$

тогда по равенству Колмогорова-Чэпмена,

$\mathbf{p}_n = (p_n(1), p_n(2) \ldots p_n(M)) = \mathbf{p}_0 \cdot \Pi^n$, где $\mathbf{p}_0 = (p_0(1), p_0(2) \ldots p_0(M))$.

Из эргодичности ДИСП следует регулярность цепи Маркова, т.е. существование предела $\Pi_\infty = \lim_{n\to\infty} \Pi^n$. Тогда все строки матрицы Π_∞ равны предельному распределению $\mathbf{p}_\infty = (p_\infty(1), p_\infty(2) \ldots p_\infty(M))$, являющемуся собственным вектором: $\mathbf{p}_\infty = \mathbf{p}_\infty \Pi$.

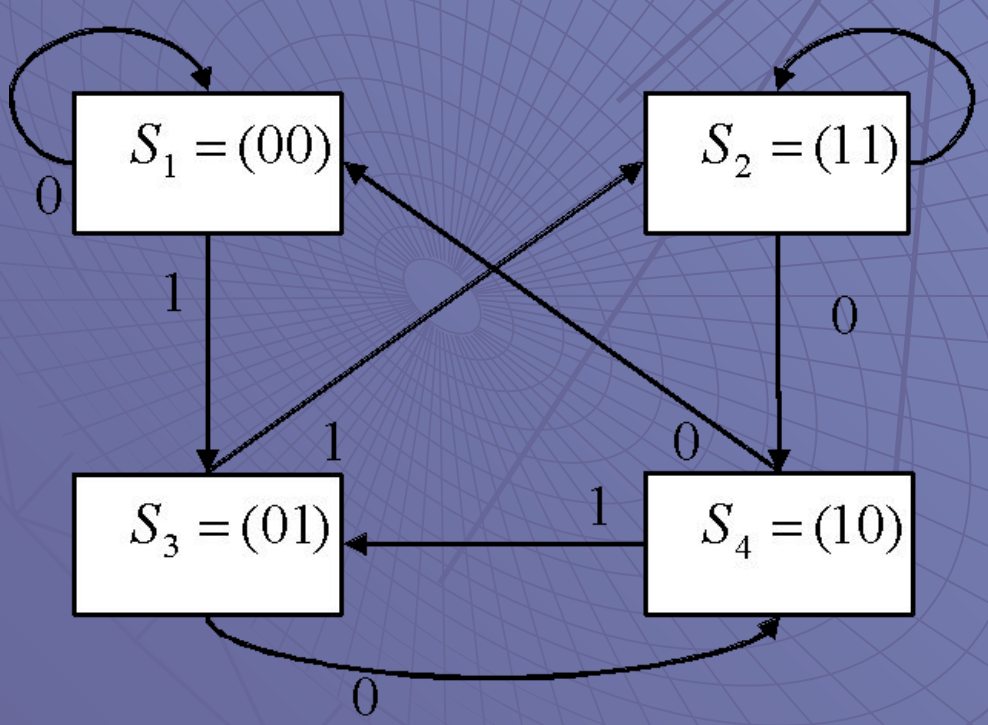
Энтропия в этом случае равна $H_\infty(X) = \sum_{j=1}^{M} p_\infty(j) H(X \mid S_j)$, где

$$H(X \mid S_j) = -\sum_{i=1}^{N} p^{(j)}(i) \cdot \log p^{(j)}(i).$$

Пример. Двоичный источник имеет память $r = 2$ и вероятности появления символов для каждого состояния даны в таблице:

$S_1 = (00)$	0	1	$S_2 = (11)$	0	1
$p^{(1)}$	0	1	$p^{(2)}$	$\frac{1}{2}$	$\frac{1}{2}$
$S_3 = (01)$	0	1	$S_4 = (10)$	0	1
$p^{(3)}$	$\frac{2}{3}$	$\frac{1}{3}$	$p^{(4)}$	$\frac{3}{4}$	$\frac{1}{4}$

Граф состояний:



Матрица переходных вероятностей:

$$\mathbf{\Pi} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{3}{4} & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

Предельная матрица $\Pi_\infty = \lim_{n\to\infty} \Pi^n = \frac{1}{41}\begin{pmatrix} 9 & 8 & 12 & 12 \\ 9 & 8 & 12 & 12 \\ 9 & 8 & 12 & 12 \\ 9 & 8 & 12 & 12 \end{pmatrix}$, значит

$\mathbf{p}_\infty = \left(\frac{9}{41} \quad \frac{8}{41} \quad \frac{12}{41} \quad \frac{12}{41}\right)$. Энтропии для каждого состояния:

$H(X \mid S_1) = 0;\ H(X \mid S_2) = 1;\ H(X \mid S_3) = 0{,}918;\ H(X \mid S_4) = 0{,}811.$

Энтропия всего источника

$$H_\infty(X) = 0 \cdot \frac{9}{41} + 1 \cdot \frac{8}{41} + 0{,}918 \cdot \frac{12}{41} + 0{,}811 \cdot \frac{12}{41} = 0{,}701.$$

Если не принимать во внимание различие состояний, то есть считать этот источник источником без памяти, то

$$P(0) = 0 \cdot \frac{9}{41} + \frac{1}{2} \cdot \frac{8}{41} + \frac{2}{3} \cdot \frac{12}{41} + \frac{3}{4} \cdot \frac{12}{41} = \frac{21}{41};$$

$P(1) = 1 \cdot \frac{9}{41} + \frac{1}{2} \cdot \frac{8}{41} + \frac{1}{3} \cdot \frac{12}{41} + \frac{1}{4} \cdot \frac{12}{41} = \frac{20}{41}$ и $H(X) = 0{,}999$.

Информационные характеристики ДИСП:

Объем памяти r; Энтропия $H_\infty(X) = \sum_{j=1}^{M} p_\infty(j) H(X \mid S_j)$;

Максимальная энтропия $H_{\max} = \log N$;

Избыточность $D = H_{\max} - H_\infty(X)$;

Относительная избыточность $D_0 = \frac{D}{H_{\max}} = \frac{H_{\max} - H_\infty(X)}{H_{\max}}$;

Средняя длительность выдачи одного символа $\tau_{cp} = \sum_{i=1}^{N} p_i \, \tau_i$,

где $p_i = \sum_{j=1}^{M} p_\infty(j) p^{(j)}(i)$;

Производительность источника $\bar{I}(X) = \frac{H_\infty(X)}{\tau_{cp}}$.