Расчетное задание 1

Идентификация сообщений, передаваемых по зашумленному каналу связи.

Дано

По каналу связи передаются буквы $[x_1;x_2;...;x_n]$ в двоичном коде. Последовательность переданных букв образует сообщение. Канал симметричный, вероятность искажения каждого отдельного символа (бита) равна q. В результате однократной передачи сообщения $X=[x^{(1)},x^{(2)}...,x^{(k)}]$ на приемной стороне принято сообщение $Y_1=[y_1^{(1)}y_1^{(2)}...y_1^{(k)}]$. В результате повторной передачи того же слова на приемной стороне принято слово $Y_2=[y_2^{(1)}y_2^{(2)}...y_2^{(k)}]$. В результате последней (m-й) передачи того же слова на приемной стороне принято слово $Y_m=[y_m^{(1)}y_m^{(2)}...y_m^{(k)}]$.

Варианты

Передаваемые буквы (алфавит) и их код приведен в табл.1 приложения. Для каждого студента есть свой вариант в виде текстового файла, названного по фамилии и имени студента. В файле задается число букв в сообщении, разрядность кода (количество бит, используемых при передаче одной буквы), шум (вероятность искажения q), число посылок m и набор из m посылок (принятых сообщений $Y^{(i)}$).

Задание

Часть 1. Последовательная передача одинаковых сообщений

1.1. Определение переданного сообщения

- вычислите априорное распределение вероятностей исходных букв алфавита p(xi), рассмотрите два случая (все дальнейшие расчеты в п. 1.1 и 1.2 необходимо будет проделать для этих двух вариантов):
  - все символы равновероятны;
  - вероятности букв задаются исходя из известной информации о частоте букв в русском алфавите (таблица 2);
- вычислите апостериорное распределение вероятностей после 1-й, 2-й и m-й передач для каждой s буквы сообщения - $P(x_i / y_1^{(s)})$, $P(x_i / y_1^{(s)} y_2^{(s)})$, $P(x_i / y_1^{(s)} y_2^{(s)} y_3^{(s)})$, … , $P(x_i / y_1^{(s)} y_2^{(s)} ..., y_m^{(s)})$; при расчете используются формулы (3), (4) и (6); следует учитывать, что для повторных посылок априорные вероятности будут совпадать с апостериорными для предыдущей посылки (см. формулу (6)).
- постройте график изменения апостериорного распределения вероятностей на примере любой l-ой передаваемой буквы сообщения (n передач => n графиков друг под другом, на графике по оси X – номер символа, по оси Y – вероятность)
- по максимуму апостериорной вероятности определите наиболее вероятные буквы и составьте вариант исходного переданного сообщения для 1-й, 2-й и m-й посылок;
- проанализируйте, как повторные передачи сказались на принятии решения.

1.2. Расчет энтропии и количества информации

- Выберите в посылаемом сообщении произвольную букву (под номером s), далее все вычисления будут относиться к этой букве;
- Определите апостериорные вероятности, рассматривая каждую передачу независимо от другой; схема вычислений следующая $P(x_i) \to P(y_j / x_i) \to P(y_j) \to P(x_i / y_j)$; при расчете используйте формулы (3), (4) и (5).
- Определите условные энтропии $H(X / y_j)$ на сообщения y_j по формуле (2), среднее количество информации $I(X, y_j)$ об X, содержащееся в y_j по формуле (8).
- Определите среднюю условную энтропию H(X/Y) по формуле (7) и среднюю взаимную информацию I(X,Y) по формуле (9).
- Постройте графики изменения условной энтропии $H(X / y_j)$ и количества информации $I(X, y_j)$ от номера посылки.

1.3. Сравните результаты п. 1.1 и 1.2 при различных заданиях изначальных априорных вероятностей.

Часть 2 Передача сообщения путем многократного дублирования

Рассмотрите m передач сообщений как передачу одного большого сообщения, в котором каждый символ многократно (m-кратно) дублируется

На входе $X_{new} = [x_{new}^{(1)}; x_{new}^{(2)};..., x_{new}^{(k)}] = [x^{(1)}x^{(1)}...x^{(1)}, x^{(2)}x^{(2)}...x^{(2)}..., x^{(k)}x^{(k)}...x^{(k)}]$,

На выходе $Y_{new} = [y_{new}^{(1)}; y_{new}^{(2)};...; y_{new}^{(k)}] = [y_1^{(1)}y_2^{(1)}...y_m^{(1)}y_1^{(2)}y_2^{(2)}...y_m^{(2)}...y_1^{(k)}y_2^{(k)}...y_m^{(k)}]$.

При этом новый алфавит по сути – m-кратное дублирование старого алфавита: $[x_{new1}; x_{new2};...; x_{newn}] = [x_1x_1...x_1; x_2x_2...x_2;...; x_nx_n...x_n]$

2.1. Определение переданного сообщения

- вычислите априорное распределение вероятностей исходных букв алфавита p(xi) – рассмотрите два случая (по аналогии с п.1. все дальнейшие расчеты в п. 2.1 и 2.2 необходимо выполнить для этих двух вариантов):
    - все символы равновероятны;
    - вероятности букв задаются исходя из известной информации о частоте букв в русском алфавите (таблица 2);
- вычислите апостериорное распределение вероятностей для каждой l буквы сообщения - $P(x_{newi} / y_{new}^{(l)})$; при расчете используются формулы (3), (4);
- постройте график апостериорного распределения вероятностей на примере l-ой передаваемой буквы сообщения
- по максимуму апостериорной вероятности определите наиболее вероятные буквы и составьте вариант исходного переданного сообщения – сравните его со случаем передачи сообщений последовательно

2.2. Расчет энтропии и количества информации

- Выберите в посылаемом сообщении ту же букву, что и использовалась в п. 1.2, далее все вычисления будут относиться к этой букве;
- Определите апостериорные вероятности; схема вычислений следующая $P(x_{newi}) \to P(y_{new} / x_{newi}) \to P(y_{new}) \to P(x_{newi} / y_{new})$; при расчете используйте формулы (3), (4) и (5).
- Определите условную энтропию $H(X_{new} / y_{new})$ на сообщения y_{new} по формуле (2), среднее количество информации $I(X, y_{new})$ об X, содержащееся в y_{new} по формуле (8).
- Определите среднюю условную энтропию $H(X_{new} / Y_{new})$ по формуле (7) и среднюю взаимную информацию $I(X, Y_{new})$ по формуле (9).
- Сравните результаты (энтропия, количество информации) с п.1.2 и объясните их.

Приложение 1 Теоретические основы

Пусть ***X*** - ансамбль возможных сообщений, которые могут быть переданы по каналам связи. Априорные вероятности, с которыми на передающей стороне может появиться **i** - ое сообщение, обозначим через $p_i = p(x_i)$.

В качестве характеристики априорной неопределенности генерации того или иного сообщения на передающей стороне Шеннон предложил применить следующий функционал энтропии:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \qquad (1)$$

В этом функционале используется логарифм по основанию **2** в связи с тем, что в цифровых каналах связи информация обычно представляется в двоичных кодах, и энтропия измеряется в *битах*.

Энтропия является характеристикой неопределенности состояния ансамбля ***X*** в отличие от характеристики возможности наступления того или иного события, то есть вероятности. Энтропия не зависит от значений, которые может принимать тот или иной элемент ансамбля и от способа его представления. В частности, элементами ансамбля могут быть словесные описания или фотографические изображения. Энтропия характеризует степень хаотичности ансамбля и принимает максимальное значение, когда вероятности $p(x_i)$ одинаковы.

Инструментарий теории информации, выдвинутой К.Шенноном, эффективно используется в теории и практике передачи информации и кодирования. В этих применениях ансамбль ***X*** - ансамбль возможных сообщений. Основные результаты теории информации относятся к передаче информации в двоичном коде и к каналам, ориентированным на передачу двоичных кодов. Передача такой информации осуществляется двоичными кодовыми словами, и при каждой такой передаче искажение отдельного символа в передаваемом слове заключается в том, что вместо **1** получатель принимает **0**, или вместо нуля получатель принимает **1**. При таких условиях канал передачи двоичной информации характеризуется вероятностью искажения символа. Схематическое представление простейшего канала связи, а именно , двоичного симметричного канала представлено на рис 5.

Симметричным этот канал называется потому, что вероятности искажения символов **'0'** и **'1'** одинаковы и равны **q**, как это показано на рисунке. Вероятность **p** - вероятность неискаженной передачи символа.

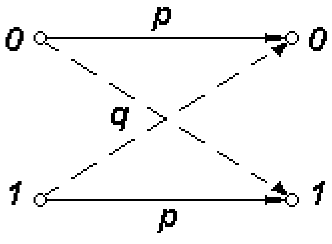

Рис.5. Двоичный симметричный канал

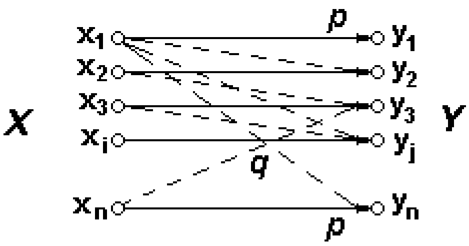

Рис.6. Передача сообщений по каналу с искажениями

Поскольку каждое сообщение $\mathbf{x}_i \in X$ передается кодовым словом, которое может состоять из нескольких символов **'0'** или **'1'**, возможностей искажения слова при передаче по такому каналу гораздо больше (см. рис. 6, где принимаемые сообщения обозначены через $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_j, \ldots \in$ ***Y***). Вероятность искажения слова в общем случае отличается от вероятности искажения отдельного символа.

Понятно, что при передаче по каналам связи сообщения (слова) $\mathbf{x_i} \in X$ и при получении сообщения $\mathbf{y_j} \in \boldsymbol{Y}$, несмотря на искажения в каналах энтропия ансамбля $\boldsymbol{X}$, несомненно, уменьшится и будет исчисляться условной энтропией на сообщение

$$H(X/y_j) = -\sum_{i=1}^{n} p(x_i/y_j) \cdot \log_2 p(x_i/y_j). \qquad (2)$$

Прим. Вероятности $p(x_i/y_j)$ называются апостериорными. Для их расчета необходимо использовать формулу Байеса

$$p(x_i/y_j) = \frac{p(y_j/x_i)p(x_i)}{p(y_j)} = \frac{p(y_j/x_i)p(x_i)}{\sum_k p(y_j/x_k)p(x_k)} \qquad (3)$$

$p(x_i)$ - априорные вероятности, вначале они могут быть выбраны произвольно и как правило они принимаются одинаковыми и равными 1/N, где N – общее количество слов xi.

$p(y_j/x_i)$ - условная вероятность приема y_j при условии, что было послано x_i. Данную вероятность легко найти следующим образом. Если общее количество разрядов в слове k, в t разрядах произошла ошибка (они проинвертировались), то вероятность очевидно равна

$$p(y_j/x_i) = p^{k-t}q^t \qquad (4)$$

Здесь по-прежнему p – вероятность правильной передачи одного бита, q=1-p – вероятность ошибки.

$$p(y_j) = \sum_k p(y_j/x_k)p(x_k) \qquad (5)$$

- вероятность приема y_j в данной задаче несет вспомогательный характер – она потребуется в том числе в дальнейших расчетах.

В случае, если при приеме одной и той же буквы несколько раз информация накапливается, для расчета апостериорных вероятностей последующих приемов в качестве априорных вероятностей необходимо использовать апостериорные вероятности после предыдущего принятого сообщения, т.е.:

$$p(x_i/y_1y_2...y_j) = \frac{p(y_j/x_i)p(x_i/y_1y_2...y_{j-1})}{p(y_j)} = \frac{p(y_j/x_i)p(x_i/y_1y_2...y_{j-1})}{\sum_k p(y_j/x_k)p(x_k/y_1y_2...y_{j-1})} \qquad (6)$$

Для того, чтобы характеризовать всю систему приема - передачи, используют среднюю условную энтропию:

$$H(X/Y) = \sum_{j=1}^{n} p(y_j) \cdot H(X/y_j) = H(X,Y) - H(Y) \qquad (7)$$

где $H(Y) = \sum_{j=1}^{n} p(y_j) \cdot \log_2 p(y_j)$, $H(X,Y) = -\sum_{j=1}^{n}\sum_{i=1}^{n} p(x_i,y_j) \cdot \log_2 p(x_i,y_j)$

$H(X,Y)$ - совместная энтропия двух ансамблей, равная

$$H(X,Y) = H(X/Y) + H(Y) = H(Y/X) + H(X)$$

Количество информации об $\mathbf{x_i}$, полученное в одном сообщении $\mathbf{y_j}$:

$$I(x_i : y_j) = \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}.$$

Среднее количество информации об $\boldsymbol{X}$, полученное в сообщении $\mathbf{y_j}$

$$I(X : y_j) = -\sum_{i=1}^{n} p(x_i/y_j) \log p(x_i) - H(X/y_j) \qquad (8)$$

Количество информации измеряется в тех же единицах, что и энтропия - в битах. Средняя взаимная информация, содержащаяся в $\boldsymbol{Y}$ об $\boldsymbol{X}$ или в $\boldsymbol{X}$ об $\boldsymbol{Y}$:

$$I(X:Y)=\sum_{i=1}^{n} p(y_j) I(X:y_j) = H(X) - H(X/Y) \qquad (9)$$

то есть это количество информации численно равно количеству неопределенности, устраненной при получении одного сообщения.

Приложение 2 Вспомогательные материалы и таблицы

Используемый набор букв и символов:
'0123456789АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЬЫЪЭЮЯабвгдеёжзийклмнопрстуфхцчшщьыъэюя.,!:?-_№() '

Каждая буква кодируется своим номером в наборе, представленным в двоичном коде (см. табл.). При передаче без ошибки буквы с кодом x7x6x5x4x3x2x1x0 на принимающей стороне будет получен такой же код x7x6x5x4x3x2x1x0.

Табл. 1 Символы и их коды

Символ	Код	Символ	Код	Символ	Код	Символ	Код	Символ	Код
0	0000000	Й	0010100	Э	0101000	р	0111100	?	1010000
1	0000001	К	0010101	Ю	0101001	с	0111101	-	1010001
2	0000010	Л	0010110	Я	0101010	т	0111110	_	1010010
3	0000011	М	0010111	а	0101011	у	0111111	№	1010011
4	0000100	Н	0011000	б	0101100	ф	1000000	(	1010100
5	0000101	О	0011001	в	0101101	х	1000001	)	1010101
6	0000110	П	0011010	г	0101110	ц	1000010	Пробел	1010110
7	0000111	Р	0011011	д	0101111	ч	1000011		
8	0001000	С	0011100	е	0110000	ш	1000100		
9	0001001	Т	0011101	ё	0110001	щ	1000101		
А	0001010	У	0011110	ж	0110010	ь	1000110		
Б	0001011	Ф	0011111	з	0110011	ы	1000111		
В	0001100	Х	0100000	и	0110100	ъ	1001000		
Г	0001101	Ц	0100001	й	0110101	э	1001001		
Д	0001110	Ч	0100010	к	0110110	ю	1001010		
Е	0001111	Ш	0100011	л	0110111	я	1001011		
Ё	0010000	Щ	0100100	м	0111000	.	1001100		
Ж	0010001	Ь	0100101	н	0111001	,	1001101		
З	0010010	Ы	0100110	о	0111010	!	1001110		
И	0010011	Ъ	0100111	п	0111011	:	1001111		

Табл. 2 Частота букв в русском языке

Буква	Частота	Буква	Частота	Буква	Частота
а	8.66	л	4.32	ц	0.52
б	1.51	м	3.29	ч	1.27
в	4.19	н	6.35	ш	0.77
г	1.41	о	9.28	щ	0.49
д	2.56	п	3.35	ъ	0.04
е	8.10	р	5.53	ы	2.11
ж	0.78	с	5.45	ь	1.90
з	1.81	т	6.30	э	0.17
и	7.45	у	2.90	ю	1.03
й	1.31	ф	0.40	я	2.22
к	3.47	х	0.92		