Министерство образования и науки РФ
Федеральное государственное бюджетное образовательное учреждение высшего образования «Томский государственный университет систем управления и радиоэлектроники» (ТУСУР)

Теория информации

Методические указания по выполнению практических работ и заданий самостоятельной подготовки для студентов ВУЗа

Томск
2018

РАССМОТРЕНО И УТВЕРЖДЕНО на заседании кафедры экономической математики, информатики и статистики факультета вычислительных систем ТУСУР.
Протокол № 10 от «23» апреля 2018 г.

Методические указания направлены на овладение основными положениями теории информации и кодирования, такими, как понятие об энтропии и количественных мерах измерения информации, основными теоремами теории информации для дискретных каналов связи, сведениями о принципах оптимального и помехоустойчивого кодирования.

Составитель:
Старший преподаватель кафедры ЭМИС Матолыгин А.А.

СОДЕРЖАНИЕ

1 Дисциплина «Теория информации»

1.1 . Цели преподавания дисциплины

Целью данного учебного курса является овладение основными положениями теории информации и кодирования

1.2 Задачи изучения дисциплины

Основной задачей изучения дисциплины является приобретение студентами прочных знаний и практических навыков в области, определяемой основной целью курса, в частности:

- овладение основными положениями теории информации и кодирования, такими, как понятие об энтропии и количественных мерах измерения информации;
- основными теоремами теории информации для дискретных каналов связи;
- сведениями о принципах оптимального и помехоустойчивого кодирования.

1.3 Требования к результатам освоения дисциплины

В результате освоения содержания дисциплины «Теория информации» студент должен:

Знать	правила определения и основные свойства энтропии непрерывных и дискретных случайных систем среднего количества информации, переносимого одним символом теоремы о пропускной способности дискретных каналов с помехами и без помех принципы оптимального и помехоустойчивого кодирования;
Уметь	рассчитывать энтропию простейших дискретных случайных систем пропускную способность дискретного канала с помехами и без помех кодировать простейшие сообщения по методу Шеннона-Фано, Хаффмена и Хемминга;
Владеть	методами расчета энтропии простейших дискретных случайных систем рассчитывать пропускную способность дискретного канала с помехами и без помех проводить кодирование простейших сообщений по методу Шеннона-Фано, Хаффмена и Хемминга.

2 Практическая работа №1 «Энтропия и ее свойства»

Цель работы. Изучить понятие информации, энтропии и научиться измерять количество информации.

Сведения являющиеся объектом хранения, передачи и преобразования, называются информацией. Мера измерения количества информации основана на понятии энтропии. Энтропия опыта из n-равновероятных исходов равна: $F(n) = \log_2 n$. Неопределенность, вносимая одним исходом составляет

$$\frac{1}{n}\log_2 n = -\frac{1}{n}\log_2\frac{1}{n} = -p\log_2 p$$

Пусть $H = -p\log_2 p$. Обобщая это выражение на ситуацию, когда опыт α имеет n неравновероятных исходов $A_1, A_2, \ldots A_n$, получим:

$$H(\alpha) = -\sum_{i=1}^{n} p(A_i)\log_2 p(A_i) \qquad (2.1)$$

Эта величина называется энтропией опыта α или энтропия по Шеннону. Энтропия является мерой неопределенности опыта, в котором проявляются случайные события, и равна средней неопределенности всех возможных его исходов. В формуле (2.1) основание логарифма может быть двоичным, десятичным или натуральным. Если используется двоичное основание, то оно может быть опущено. Энтропия характеризует среднее значение и представляет собой математическое ожидание от $log\ p$.

Свойства энтропии:

1.Энтропия есть величина вещественная, ограниченная и неотрицательная.

2.Энтропия минимальна и равна нулю, если хотя бы одно из состояний системы достоверно известно.

3. Энтропия максимальна и равна логарифму числа состояний, если состояния системы равновероятны: $H=H_{max}= log\ n$.

4. Энтропия бинарных величин изменяется от 0 до 1.

Задание

Рассмотрите третье свойство энтропии. Будем считать, что имеется три системы, каждая из которых имеет по четыре состояния. Распределения вероятностей задает для каждого студента преподаватель, например:

система 1 0,5; 0,2; 0,25; 0,05;

система 2 0,25; 0,25; 0,25; 0,25;

система 3 0,2; 0,3, 0,28, 0,22.

Определите распределения вероятностей при которых энтропия максимальна.

Вопросы для самопроверки:

1. Почему в определении энтропии как меры неопределенности выбрана логарифмическая зависимость между H и n? Почему выбран log2 ?

2. Вопрос имеет два варианта ответа. Возможно ли, чтобы с каждым из ответов была связано различное количество информации?

3. Возможно ли, чтобы бинарный ответ содержал меньше 1 бит информации?

4. Какое количество информации содержит каждый из ответов на вопрос, если всего их 3 и все они равновероятны? А если равновероятных ответов n?

3 Практическая работа №2 «Условная энтропия и ее свойства»

Цель работы. Изучить понятие условной энтропии.

Пусть имеются две зависимые системы X и Y. Обозначим условную вероятность $\mathrm{p}(y_j|x_i)$ того, что система Y примет состояние y_j при условии, что система X приняла x_i.

Определим условную частную энтропию системы Y относительно отдельного события x_i. При этом должны быть известны условные вероятности $\mathrm{p}(y_j|x_i)$/ Тогда частная условная энтропия будет равна

$$H(Y|x_i) = -\sum_{i=1}^{n} \mathrm{p}(y_j|x_i)\, log_2 \mathrm{p}(y_j|x_i).$$

Частную условную энтропию можно выразить через математическое ожидание

$$H(Y|x_i) = M[-log_2 \mathrm{p}(Y|x_i)].$$

Чтобы полностью охарактеризовать энтропию системы, нужно определить полную или среднюю энтропию. Если частную, условную энтропию усреднить по всем состояниям x_i с учетом вероятности появления каждого из состояний $\mathrm{p}(x_i)$, то найдем полную условную энтропию сообщений Y относительно X.

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{p}(x_i, y_j)\, log_2 \mathrm{p}(y_j|x_i).$$

Понятие условной энтропии широко используется для определения информационных потерь при передаче информации. Пусть по каналу связи передаются сообщения с помощью алфавита X. В результате воздействия помех приемником будет восприниматься другой алфавит Y. $\mathrm{H}(y_j|x_i)$ выражает неопределенность того что, отправив x_i, мы получим y_j, а понятие $\mathrm{H}(x_j|y_i)$ – неуверенность, которая остается после получения y_j в том, что было отправлено именно x_i. Если в канале связи помехи отсутствуют, то всегда посланному символу x_1 соответствует принятый символ y_1, $x_2 \to y_2$, . . ., $x_n \to y_n$. При этом энтропия источника H(X) равна энтропии приемника H(Y). Если в канале связи присутствуют помехи, то они уничтожают часть передаваемой информации.

Для вычисления энтропии совместного появления статистически зависимых сообщений используется понятие энтропии объединения. Энтропия объединения представляет собой сумму вида:

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

Энтропия объединения и условная энтропия связаны между собой соотношением:

$$\mathrm{H(X,Y) = H(X) + H(X/Y) = H(Y) + H(X/Y)}$$

В случае если X и Y между собой независимы, то условная энтропия равняется безусловной.

В общем случае энтропия объединенной системы H(X,Y)≤ H(X)+H(Y), это следует из того что: H(Y/X)≤H(Y) условная энтропия меньше безусловной. Энтропия объединенной системы достигает максимум, только в случае если системы независимы.

В случае полной зависимости систем, состояния одной системы полностью определяют состояния другой (они эквивалентны):

H(X,Y)=H(X)=H(Y), так как H(Y/X)=0.

Основные свойства условной энтропии.

1. Если сообщения X и Y статистически независимы, то условная энтропия сообщений Y относительно X равна безусловной энтропии сообщений Y: H(Y/X)=H(Y). В этом случае вся информация, которую содержат сообщения Y, является новой по отношению к информации, содержащейся в сообщениях X.

2. Если сообщения X и Y являются статистически жестко связанными, то условная энтропия сообщений Y относительно X равна нулю H(Y/X)=0. Это означает, что сообщения Y не содержат никакой новой информации сверх той, которая содержится в сообщениях X.

3. Условная энтропия всегда меньше безусловной энтропии H(Y/X)<H(Y).

Задание

Рассмотрите третье свойство энтропии. Будем считать, что имеется три системы, каждая из которых имеет по четыре состояния. Распределения вероятностей задает для каждого студента преподаватель, например:

система 1 0,5; 0,2; 0,25; 0,05;

система 2 0,25; 0,25; 0,25; 0,25;

система 3 0,2; 0,3, 0,28, 0,22.

Определите распределения вероятностей при которых энтропия максимальна.

Вопросы для самопроверки:

1. Дайте определение условной энтропии.

2. Перечислите основные свойства условной энтропии.

4 Практическая работа №3 «Количество информации как мера снятой неопределенности»

Цель работы. Изучить понятие условной энтропии.

При передаче сообщений, о какой либо системе происходит уменьшение неопределенности. Если о системе все известно, то нет смысла посылать сообщение. Например, если мы получим сообщение, что Москва столица России известно. Но если мы получаем сведения о системе для нас неизвестной, то получим соответствующее количество информации. И чем более неопределенным было состояние системы, тем большее количество информации мы получим. Поэтому количество информации измеряют уменьшением энтропии.

Рассмотрим систему X над которой производится наблюдение. Допустим, что мы получили полную информацию о системе X. Пусть до наблюдения энтропия системы была равна H. В результате наблюдения энтропия равна нулю H=0, то есть все о системе известно. Тогда информация будет равна уменьшению энтропии: I=H-0, I=H.

То есть, количество информации, приобретаемое при полном выяснении состояния некоторой физической системы, равно энтропии этой системы. Количество информации I есть осредненное значение логарифма вероятности состояния. Полная информация будет представлять математическое ожидание от частных значений I. Если все возможные состояния системы априори одинаково вероятны, то частная информация равна средней (полной) информации.

В случае если состояния системы обладают различными вероятностями, информация от разных сообщений неодинакова: наибольшую информацию несут сообщения о тех событиях, которые априори были наименее вероятны. Например, сообщение о том, что 31 декабря в Томске выпал снег, несет гораздо меньше информации, чем аналогичное сообщение, что 1 июля выпал снег.

Задание

Вычислите количество информации для русского алфавита с помощью MS Excel по данным, приведенным в таблице.

Буква	**пробел**	**о**	**е, ё**	**а**	**и**	**т**	**н**	**с**
Относительная частота	0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
Буква	**р**	**в**	**л**	**к**	**м**	**д**	**п**	**у**
Относительная частота	0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
Буква	**я**	**ы**	**з**	**ь, ъ**	**б**	**г**	**ч**	**й**
Относительная частота	0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
Буква	**х**	**ж**	**ю**	**ш**	**ц**	**щ**	**э**	**ф**
Относительная частота	0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

Вопросы для самопроверки:

1. Как зависит неопределенность системы от полученной информации?

2. Допустим, что нам интересно знать, сдал или не сдал экзамен некий студент Иванов. Примем следующие вероятности этих двух событий: p(сдал)=6/8, p(не сдал)=2/8. Отсюда видно, что этот студент является довольно сильным. Приведите пример сообщения, которое несет мало информации.

3. Тексты, составленные из 32 букв алфавита, передаются по телетайпу при помощи двух качественных признаков (1 и 0). Чему равно количество информации, приходящееся на одну принятую букву, на k принятых букв?

5 Практическая работа №4 «Кодирование. Пропускная способность канала»

Цель работы. Изучить понятие пропускной способности канала.

Одной из важных характеристик источника является скорость создания информации. Для стационарного дискретного источника она равна энтропии последовательности символов на его выходе. В качестве скорости создания информации стационарным непрерывным источником принимают значение энтропии непрерывной функции на его выходе, вычисленной при заданной точности воспроизведения этой функции. Скоростью передачи информации называется средняя величина взаимной информации (в единицу времени или на отсчет) между сигналом X(t) на выходе источника и сигналом Y(t), поступающим к потребителю.

Скорость передачи информации зависит в значительной степени от скорости её создания, способов кодирования и декодирования. Наибольшая возможная в данном канале скорость передачи информации называется его пропускной способностью. Пропускная способность канала, по определению, есть скорость передачи информации при использовании «наилучших» для данного канала источника, кодера и декодера, поэтому она характеризует только канал. Пропускная способность дискретного (цифрового) канала без помех $C=log\ m$. Скорость передачи информации в дискретном канале без шумов равна его пропускной способности, когда символы в канале независимы, а все m букв алфавита равновероятны (используются одинаково часто).

Задание

Канал связи описан канальной матрицей P(Y| X) . Вычислите среднее количество информации, которое переносится одним символом сообщения, если вероятности появления символов источника сообщений равны $p(x_1)$; $p(x_2)$; $p(x_3)$. Чему равны информационные потери при передаче сообщения из 1000 символов алфавита x_1, x_2, x_3? Чему равно количество принятой информации? Канальная матрица и вероятности появления символов источника сообщений для каждого студента задаются индивидуально.

Вопросы для самопроверки:

1. Как определить скорость создания информации стационарным непрерывным источником?

2. От чего зависит скорость передачи информации?

6 Практическая работа №5 «Эффективное кодирование»

Цель работы. Освоить основы эффективного кодирования.

При сжатии и передаче информации по каналам с шумом (помехами) используются следующие виды кодирования: эффективное кодирование. Эффективное кодирование базируется на основной теореме Шеннона для каналов без шума, который доказал, что сообщения, составленные из букв некоторого алфавита, можно закодировать так, что среднее число двоичных символов на букву будет сколь угодно близко к энтропии источника этих сообщений, но не меньше этой величины.

Эффективное кодирование заключается в том, что, учитывая статистические свойства какого-либо сообщения, можно минимизировать среднее число двоичных символов, требующихся для выражения одной буквы сообщения, что при отсутствии шума позволяет уменьшить время передачи или объем запоминающего устройства. Используется в архиваторах для сжатия информации. К числу эффективных кодов относятся коды: Шеннона-Фэно, Хаффмена и другие.

Задание

Постройте эффективный код для сообщения, заданного преподавателем. Для построения кода воспользуйтесь методикой Шеннона-Фэно. По этой методике код строится следующим образом. Буквы алфавита сообщений выписываются в таблицу в порядке убывания вероятностей. Затем они разделяются на две группы так, чтобы суммы вероятностей в каждой из групп были по возможности одинаковы. Всем буквам верхней половины в качестве первого символа приписывается 1, а всем нижним – 0. Каждая из полученных групп разбивается на две подгруппы с одинаковыми суммарными вероятностями и т. д. Процесс повторяется до тех пор, пока в каждой подгруппе останется по одной букве.

Вопросы для самопроверки:

1. В каких случаях применяют эффективное кодирование?
2. В чем заключается эффективное кодирование?

7 Практическая работа №6 «Код Хемминга»

Цель работы. Освоить основы кодирования методом Хемминга.

Коды Хэмминга являются самоконтролирующимися кодами, то есть кодами, позволяющими автоматически обнаруживать ошибки при передаче данных. Для их построения достаточно приписать к каждому слову один добавочный (контрольный) двоичный разряд и выбрать цифру этого разряда так, чтобы общее количество единиц в изображении любого числа было, например, четным. Одиночная ошибка в каком-либо разряде передаваемого слова (в том числе, может быть, и в контрольном разряде) изменит четность общего количества единиц. Счетчики по модулю 2, подсчитывающие количество единиц, которые содержатся среди двоичных цифр числа, могут давать сигнал о наличии ошибок.

При этом невозможно узнать, в каком именно разряде произошла ошибка, и, следовательно, нет возможности исправить её. Остаются незамеченными также ошибки, возникающие одновременно в двух, в четырёх или вообще в четном количестве разрядов. Впрочем, двойные, а тем более четырёхкратные ошибки полагаются маловероятными.

Задание

Работа выполняется на персональном компьютере с использованием языков программирования высокого уровня. Постройте программу, которая бы позволила ввести сообщение произвольной длины. Программа должна выделить буквы алфавита введенного текста, подсчитать и выдать частоту появления этих букв. Определить энтропию, приходящуюся в среднем на одну букву, длину кода при равномерном кодировании и избыточность.

Используя результаты, выданные компьютером:

1) проанализируйте алфавит введенного сообщения (количество символов алфавита, значение энтропии H, длину кода Q при целочисленном кодировании, избыточность R при однобуквенном кодировании);

2) постройте вручную код Хемминга (используя полученные статистические данные).

Вопросы для самопроверки.

1. Как строится код Хемминга?

2. Каковы слабые стороны кода Хемминга?

8 Практическая работа №6 «Код Шеннона-Фано, код Хаффмана»

Цель работы. Приобретение практических навыков построения кода Хаффмана, кода Шеннона-Фано.

Алгоритм Шеннона –Фано это один из первых алгоритмов сжатия. Данный метод сжатия имеет большое сходство с кодом Хаффмана, который появился позже. Алгоритм использует коды переменной длины: часто встречающийся символ кодируется кодом меньшей длины, редко встречающийся – кодом большей длины. Коды Шеннона–Фано префиксные. Это свойство позволяет однозначно декодировать любую последовательность кодовых слов. Алгоритм этого кода имеет несколько этапов.

1. Символы первичного алфавита выписывают в порядке убывания вероятностей.

2. Символы полученного алфавита делят на две части, суммарные вероятности символов которых максимально близки друг другу.

3. В префиксном коде для первой части алфавита присваивается двоичная цифра «0», второй части – «1».

4. Полученные части рекурсивно делятся и их частям назначаются соответствующие двоичные цифры в префиксном коде.

Когда размер подалфавита становится равен нулю или единице, то дальнейшего удлинения префиксного кода для соответствующих ему символов первичного алфавита не происходит, таким образом, алгоритм присваивает различным символам префиксные коды разной длины. На шаге деления алфавита существует неоднозначность, так как разность суммарных вероятностей может быть одинакова для двух вариантов разделения (учитывая, что все символы первичного алфавита имеют вероятность больше нуля).

Результатом построения кода Хаффмана является список кодовых слов и список длин кодовых слов, которые необходимы для кодирования и декодирования. Работа алгоритма начинается с составления списка символов (чисел) алфавита в порядке убывания их частоты (вероятности). Результатом построения кода Хаффмана является список кодовых слов и список длин кодовых слов, необходимых для кодирования и декодирования. Код Хаффмана строят с обходом и без обхода кодового дерева.

Чтобы произвести декодирование данных необходимо знать построенные коды. Самый лучший способ передать декодеру частоты появления символов в потоке. На основе этих частот декодер строит коды, а затем применяет их для декодирования последовательности. Поэтому, чем длиннее кодированная последовательность, тем меньше удельный вес служебной информации переданной декодеру.

Задание

Построить код Хаффмана текстового файла, выданного преподавателем. Построить одномерное распределение вероятностей символов в текстовом файле. Исключить из рассмотрения символы с нулевой вероятностью. Вычислить энтропию источника информации. Вычислить среднюю длину кодового слова и избыточность кода. Реализовать кодер в соответствии с вариантом задания. Сформировать сжатый файл. Убедиться, что размер сжатого файла примерно равен In, где n - количество символов в исходном текстовом файле. Реализовать декодер в соответствии с вариантом задания. Убедиться, что исходный и восстановленный текстовые файлы полностью совпадают по размеру и по содержимому.

Вопросы для самопроверки.

1. С чего начинается построение дерева в коде Шеннона-Фано?

2. В чем отличие кода Шеннона-Фано от кода Хаффмана?

Список литературы.

1. Акулиничев Ю.П. Теория и техника передачи информации : учебное пособие / Ю. П. Акулиничев, А. С. Бернгардт. — Томск: Эль Контент, 2012. — 210 с. URL: http://edu.tusur.ru/training/publications/1750 (дата обращения 07.09.12 г.)

2. Основы теории информации: курс лекций / Ю.В. Свирид. – Мн.: БГУ. – 2003.–С. 97-107. URL: http://www.svirid.by/source/Lectures_ru.pdf

3. Теория информации и кодирование : Учебное пособие для вузов / Б. Б. Самсонов [и др.]. - Ростов н/Д : Феникс, 2002. - 288 с.