Лекция 1. Введение. Базовые понятия.

Аннотация.

Вводная информация: цели курса, его состав, требования к студентам, система оценок и прочая орг. информация.

Информация и ее виды. Меры информации, критерии качества.

Оглавление

Введение: организация занятий, цели курса, виды контрольных мероприятий.

Информация и ее виды.

[Лидовский]

Кибернетика, теория информации.

Теория информации рассматривается как существенная часть кибернетики.

Кибернетика – это наука об общих законах получения, хранения, передачи и переработки информации.

Объектом исследования кибернетики являются все управляемые системы.

Примеры кибернетических систем: автоматические регуляторы в технике, ЭВМ, мозг человека или животных, биологическая популяция, социум. Часто кибернетику связывают с методами искусственного интеллекта, т.к. она разрабатывает принципы создания систем управления и систем автоматизации умственного труда.

Основными разделами кибернетики считаются: теория информации, теория алгоритмов, теория автоматов, исследование операций, теория оптимального управления и теория распознавания образов.

Родоначальниками кибернетики считаются американские ученые Норберт Винер и Клод Шеннон. Клод Шеннон – основоположник теории информации.

Но́рберт Ви́нер (англ. Norbert Wiener; 26 ноября 1894, Колумбия, штат Миссури, США — 18 марта 1964, Стокгольм, Швеция) — американский учёный, выдающийся математик и философ, основоположник кибернетики и теории искусственного интеллекта.

В 4 года Винер уже был допущен к родительской библиотеке, а в 7 лет написал свой первый научный трактат по дарвинизму. Норберт никогда по-настоящему не учился в средней школе. Зато 11 лет от роду он поступил в престижный Тафт-колледж, который закончил с отличием уже через три года получив степень бакалавра искусств.

В 18 лет Норберт Винер получил степени доктора философии по математической логике в Корнельском и Гарвардском университетах. В девятнадцатилетнем возрасте доктор Винер был приглашён на кафедру математики Массачусетского технологического института.

Винер ввел основную категорию кибернетики – управление, показал существенные отличия этой категории от других, например, энергии, описал несколько задач, типичных для кибернетики и привлек всеобщее внимание к особой роли вычислительных машин, считая их индикатором наступления новой НТР. Выделение категории управления позволило Винеру воспользоваться понятием информации, положив в основу кибернетики изучение законов передачи и преобразования информации.

Кибернетика — наука об оптимальном управлении сложными динамическими системами, изучающая общие принципы управления и связи, лежащие в основе работы самых разнообразных по природе систем — от самонаводящих ракет-снарядов и быстродействующих вычислительных машин до сложного живого организма.

Оптимальное управление — это перевод системы в новое состояние с выполнением некоторого критерия оптимальности, например, минимизации затрат времени, труда, веществ или энергии.

Сущность принципа управления заключается в том, что движение и действие больших масс или передача и преобразование больших количеств энергии направляется и контролируется при помощи небольших количеств энергии, несущих информацию. Этот принцип управления лежит в основе организации и действия любых управляемых систем.

Теория информации тесно связана с такими разделами математики, как теория вероятностей и математическая статистика, которые представляют для нее математический фундамент. Теория информации представляет собой математическую теорию, посвященную измерению информации, ее потока, характеристик канала связи и т.п., особенно применительно к радио, телевидению и другим средствам связи. Первоначально теория была посвящена каналу связи, определяемому длиной волны электромагнитного излучения или колебаний воздуха

Базовые понятия теории информации.

Информация – нематериальная сущность, при помощи которой с любой точностью можно описывать реальные (материальные), виртуальные(возможные) и понятийные сущности. Иначе, **информация** - сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые уменьшают имеющуюся о них степень неопределенности, неполноты знаний.

Теория информации (математическая теория связи) — раздел прикладной математики, аксиоматически определяющий понятие информации[1], её свойства и устанавливающий предельные соотношения для систем передачи данных.

Использует, главным образом, математический аппарат теории вероятностей и математической статистики.

Основные разделы теории информации — кодирование источника (сжимающее кодирование) и канальное (помехоустойчивое) кодирование. Теория информации тесно связана с криптографией и другими смежными дисциплинами.

Канал связи – это среда передачи информации, которая характеризуется в первую очередь максимально возможной для нее скоростью передачи данных(емкостью канала связи).

Шум – это помехи в канале связи при передаче информации.

Кодирование – преобразование дискретной информации одним из следующих способов: шифрование, сжатие, защита от шума.

Информатика рассматривает информацию как концептуально связанные между собой сведения, данные, понятия, изменяющие наши представления о явлении или объекте окружающего мира. Наряду с информацией в информатике часто употребляется понятие *данные*. Покажем, в чем их отличие.

Данные могут рассматриваться как признаки или записанные наблюдения, которые по каким-то причинам не используются, а только хранятся. В том случае, если появляется возможность использовать эти данные для уменьшения неопределенности о чем-либо, данные превращаются в информацию.

Формы адекватности информации

Адекватность информации - это определенный уровень соответствия создаваемого с помощью полученной информации образа реальному объекту, процессу, явлению и т.п.

Адекватность информации может выражаться в трех формах: семантической, синтаксической, прагматической.

Синтаксическая адекватность. Она отображает формально-структурные характеристики информации и не затрагивает ее смыслового содержания. На синтаксическом уровне учитываются тип носителя и способ представления информации, скорость передачи и обработки, размеры кодов представления информации, надежность и точность преобразования этих кодов и т.п. Информацию, рассматриваемую только с синтаксических позиций, обычно называют данными, так как при этом не имеет значения смысловая сторона. Эта форма способствует восприятию структурных внешних характеристик, т.е. синтаксической стороны информации.

Семантическая (смысловая) адекватность. Эта форма определяет степень соответствия образа объекта и самого объекта. Семантический аспект предполагает учет смыслового содержания информации. На этом уровне анализируются те сведения, которые отражает информация, рассматриваются смысловые связи. В информатике устанавливаются смысловые связи между кодами представления информации.

Эта форма служит для формирования понятий и представлений, выявления смысла, содержания информации и ее обобщения.

Прагматическая (потребительская) адекватность. Она отражает отношение информации и ее потребителя, соответствие информации цели управления, которая на ее основе реализуется. Проявляются прагматические свойства информации только при наличии единства информации (объекта), пользователя и цели управления. Прагматический аспект рассмотрения связан с ценностью, полезностью использования информации при выработке потребителем решения для достижения своей цели.

Меры информации

Классификация мер

Для измерения информации вводятся два параметра: количество информации *I* и объем данных $V_Д$.

Эти параметры имеют разные выражения и интерпретацию в зависимости от рассматриваемой формы адекватности. Каждой форме адекватности соответствует своя мера количества информации и объема данных.

Синтаксическая мера информации

Эта мера количества информации оперирует с обезличенной информацией, не выражающей смыслового отношения к объекту.

Объем данных $V_Д$ в сообщении измеряется количеством символов (разрядов) в этом сообщении. В различных системах счисления один разряд имеет различный вес и соответственно меняется единица измерения данных: в двоичной системе счисления единица измерения - бит *(bit — binary digit*—двоичный разряд);

в десятичной системе счисления единица измерения - дит (десятичный разряд).

Количество информации I на синтаксическом уровне невозможно определить без рассмотрения понятия неопределенности состояния системы (энтропии системы).

1865 г. немецкий физик Рудольф Клазиус ввел в статистическую физику понятие энтропии как меры уравновешенности системы. В 1921 году основатель большей части математической статистики, англичанин Роналд Фишер впервые ввел термин «информация» в математику, но полученные им формулы носят очень специальный характер.

В 1948 году Клод Шеннон в своих работах по теории связи выписывает формулы для вычисления количества информации и энтропии. Термин «энтропия» используется Шенноном по совету фон Неймана, отметившего совпадение полученных Шенноном формул с соответствующими формулами статистической физики.

Информация через неопределенность

Действительно, получение информации о какой-либо системе всегда связано с изменением степени неосведомленности получателя о состоянии этой системы. Рассмотрим это понятие.

Пусть до получения информации потребитель имеет некоторые предварительные (априорные) сведения о системе α. Мерой его неосведомленности о системе является функция *H(α),* которая в то же время служит и мерой неопределенности состояния системы.

После получения некоторого сообщения β получатель приобрел некоторую дополнительную информацию $I_\beta(\alpha)$, уменьшившую его априорную неосведомленность гак, что апостериорная (после получения сообщения β) неопределенность состояния системы стала $H_\beta(\alpha)$.

Тогда количество информации $I_\beta(\alpha)$ о системе, полученной в сообщении β, определится как $I_\beta(\alpha) = H(\alpha) - H_\beta(\alpha)$, т.е. количество информации измеряется изменением (уменьшением) неопределенности состояния системы.

Если конечная неопределенность $H_\beta(\alpha)$ обратится в нуль, то первоначальное неполное знание заменится полным знанием и количество информации $I_\beta(\alpha) = H(\alpha)$.

Иными словами *энтропия системы* H(α) может рассматриваться как мера недостающей информации.

Энтропия системы H(α), имеющая *N* возможных состояний, согласно формуле Шеннона, равна:

$$H(\alpha) = -\sum_{i=1}^{N} P_i \log P_i$$

где P_i - вероятность того, что система находится в i-м состоянии.

Для случая, когда все состояния системы равновероятны, т.е. их вероятности равны,

$P_i = \frac{1}{N}$ ее энтропия определяется соотношением.

$$H(\alpha) = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N}$$

Часто информация кодируется числовыми кодами в той или иной системе счисления, особенно это актуально при представлении информации в компьютере.

Естественно, что одно и то же количество разрядов в разных системах счисления может передать разное число состояний отображаемого объекта, что можно представить в виде соотношения

$$N=m^n,$$

где *N* — число всевозможных отображаемых состояний;

m - основание системы счисления (разнообразие символов, применяемых в алфавите);

n - число разрядов (символов) в сообщении.

Наиболее часто используются двоичные и десятичные логарифмы. Единицами измерения в этих случаях будут соответственно бит и дит.

Теорема Хартли: Информативность символа m-элементного алфавита равна *log m.*

Формула Шеннона выражает информативность источника информации с *m*-символьным алфавитом и данной частотной характеристикой.

Но каналу связи передается n-разрядное сообщение, использующее *m* различных символов. Так как количество всевозможных кодовых комбинаций будет $N=m^n$, то при равновероятности появления любой из них количество информации, приобретенной абонентом в результате получения сообщения, будет $I=logN=logm$ - формула Хартли.

Если в качестве основания логарифма принять *m,* то $I=n$. В данном случае количество информации (при условии полного априорного незнания абонентом содержания сообщения) будет равно объему данных $I=V_д$, полученных по каналу связи. Для неравновероятных состояний системы всегда $I<V_д=n$.

Коэффициент (степень) информативности (лаконичность) сообщения определяется отношением количества информации к объему данных, т.е.

$$Y = \frac{I}{V_д}, \text{ причем } Y<0<1,$$

С увеличением *Y* уменьшаются объемы работы по преобразованию информации (данных) в системе. Поэтому стремятся к повышению информативности, для чего разрабатываются специальные методы оптимального кодирования информации.

Формулы Шеннона

Равномерное распределение имеет наибольшую энтропию среди всех распределений с данным числом исходов.

Предложенный Шенноном способ измерения количества информации, содержащейся в одной случайной величине относительно другой случайной величины лежит в основе теории информации. Он приводит к числовой записи количества информации.

Для ДСВ(дискретных случайных величин) X и Y, заданных законами распределения $P(X = X_i) = p_i$, $P(Y = Y_i) = q_j$ и совместным распределением $P(X = X_i, Y = Y_i) = p_{ij}$, количество информации, содержащейся в X относительно Y, равно

$$I(X,Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j}$$

Для НСВ(непрерывных случайных величин) X и Y, заданных плотностями распределения вероятностей px(t1), py(t2) и pxy(t1, t2), аналогичная формула имеет вид

$$I(X,Y) = \iint\limits_{\mathbb{R}^2} p_{XY}(t_1,t_2) \log_2 \frac{p_{XY}(t_1,t_2)}{p_X(t_1)p_Y(t_2)} dt_1 dt_2.$$

Очевидно, что

$$P(X = X_i, X = X_j) = \begin{cases} 0, & \text{при } i \neq j \\ P(X = X_i), & \text{при } i = j \end{cases}$$

и, следовательно,

$$I(X,X) = \sum_i p_i \log_2 \frac{p_i}{p_i p_i} = -\sum_i p_i \log_2 p_i.$$

Энтропия д.с.в. X в теории информации определяется формулой

$$H(X) = HX = I(X,X).$$

Свойства меры информации и энтропии:

1) $I(X,Y) \geqslant 0$, $I(X,Y) = 0 \quad \Leftrightarrow \quad X$ и Y независимы;
2) $I(X,Y) = I(Y,X)$;
3) $HX = 0 \quad \Leftrightarrow \quad X$ — константа;
4) $I(X,Y) = HX + HY - H(X,Y)$, где $H(X,Y) = -\sum_{i,j} p_{ij} \log_2 p_{ij}$;
5) $I(X,Y) \leqslant I(X,X)$. Если $I(X,Y) = I(X,X)$, то X — функция от Y.

Пример использования энтропии Шеннона.

Энтропия ДСВ – это минимум среднего количества бит, которое нужно передавать по каналу связи о текущем значении данной ДСВ.

Рассмотрим пример (скачки). В заезде участвуют 4 лошади с равными шансами на победу. Введем ДСВ X, равную номеру победившей лошади. Здесь HX = 2. После каждого заезда по каналам связи достаточно будет передавать два бита информации о номере победившей лошади. Кодируем номер лошади следующим образом: 1-00, 2-01, 3-10, 4-11. Если ввести функцию L(X), которая возвращает длину сообщения, кодирующего заданное значение X, то м.о. ML(X) – это средняя длина сообщения, кодирующего X. Можно формально определить L через две функции L(X) = len(code), где code(X) каждому значению X ставит в соответствие некоторый битовый код, причем, взаимно однозначно, а len возвращает длину в битах для любого конкретного кода. В этом примере ML(X)=HX.

Пусть теперь ДСВ X имеет следующее распределение

$$P(X=1) = \frac{3}{4}, P(X=2) = \frac{1}{8}, P(X=3) = P(X=4) = \frac{1}{16}$$

Тогда $HX = \frac{3}{4}\log_2 \frac{4}{3} + \frac{1}{8}\log_2 \frac{8}{1} + \frac{1}{8}\log_2 \frac{16}{1} = 1.186 \frac{\text{бит}}{\text{сим}}$

Закодируем номера лошадей: 1-0, 2-10, 3-110, 4-111, - т.е. так, чтобы каждый код не был префиксом другого кода (такое кодирование называется префиксным). В среднем в 16 заездах 1-я лошадь должна победить 12 из них, 2-я – в 2-х, 3-я в 1-м и 4-я – в одном. Таким образом, средняя длина сообщения о победителе равна (1*12+2*2+3*1+3*1)/16 = 1.375 бит/сим. Или м.о. L(X). Действительно, L(X) сейчас задается следующим распределением вероятностей: P(L(X)=1) = ¾, P(L(X)=2) = 1/8, P(L(X)=3) = 1/8. Следовательно,

M(L(X)) = ¾+2/8+3/8=11/8 = 1.375 бит/сим.

Таким образом, ML(X)>HX.

Может быть доказано, что более эффективного кодирования для рассмотренного случая не существует.

То, что энтропия Шеннона соответствует интуитивному представлению о мере информации, может быть продемонстрировано в опыте по определению среднего времени психических реакций. Опыт заключается в том, что перед испытуемым человеком зажигается одна из N-лампочек, которую он должен указать, Проводится большая серия испытаний, в которых каждая лампочка зажигается с определенной вероятностью. Оказывается, среднее время, необходимое для правильного ответа испытуемого, пропорционально величине энтропии ($-\sum_{i=1}^{N} p_i \log_2 p_i$), а не числу лампочек, как можно было подумать. В этом опыте предполагается, что чем больше информации будет получено человеком, тем дольше будет время ее обработки и, соответственно, реакции на нее.

Семантическая мера информации

Для измерения смыслового содержания информации, т.е. ее количества на семантическом уровне, наибольшее признание получила тезаурусная мера, которая связывает семантические свойства информации со способностью пользователя принимать поступившее сообщение. Для этого используется понятие *тезаурус пользователя.*

Тезаурус - это совокупность сведений, которыми располагает пользователь или система.

В зависимости от соотношений между смысловым содержанием информации *S* и тезаурусом пользователя S_P изменяется количество семантической информации 1_C, воспринимаемой пользователем и включаемой им в дальнейшем в свой тезаурус.

Характер такой зависимости показан на рис. 2. Рассмотрим два предельных случая, когда количество семантической информации I_C равно 0:

- при $S_P = 0$ пользователь не воспринимает, не понимает поступающую информацию;
- при $S_P \rightarrow \infty$ пользователь все знает, и поступающая информация ему не нужна.

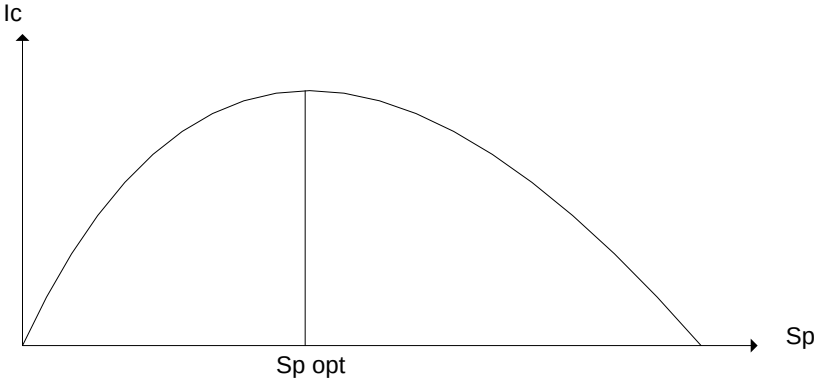

Рис. 2. Зависимость количества семантической информации, воспринимаемой потребителем, от его тезауруса $I_C = f(S_H)$

Максимальное количество семантической информации I_c потребитель приобретает при согласовании ее смыслового содержания *S* со своим тезаурусом S_P ($S_P = S_{P\,opt}$), когда поступающая информация понятна пользователю и несет ему ранее не известные (отсутствующие в его тезаурусе) сведения.

Следовательно, количество семантической информации в сообщении, количество новых знаний, получаемых пользователем, является величиной относительной. Одно и то же сообщение может иметь смысловое содержание для компетентного пользователя и быть бессмысленным (семантический шум) для пользователя некомпетентного.

При оценке содержательного аспекта информации необходимо стремиться к согласованию величин *S* и S_P.

Относительной мерой количества семантической информации может служить коэффициент содержательности С, который определяется как отношение количества семантической информации к ее объему:

$$C = \frac{I_C}{V_Д}$$

Функции меры семантической информации.

В 50-х годах XX века появились первые попытки определения абсолютного информационного содержания предложений естественного языка. Шеннон однажды заметил, что смысл сообщений не имеет никакого отношения к его теории информации, целиком построенной на положениях теории вероятностей. Но его способ точного измерения информации наводил на мысль о возможности существования способов точного измерения информации более общего вида, например, информации из

предложений естественного языка. Примером одной из таких мер является функция *inf(s)* = -$\log_2$p(s), где *s* - это предложение, смысловое содержание которого измеряется, *p(s)* - вероятность истинности s. Вот некоторые свойства этой функции-меры:

1) если s_l => s_2 (из s_1 следует s_2) — истинно, то *inf*(s_1) >= inf(s_2);
2) *inf(s)>=0;*
3) если s — истинно, то *inf(s)* = 0;
4) *inf*($s_l s_2$) = *inf*(s_l) + inf(s_2) <=> *p*(s1 s2) = *p(s1)p*(s_2), т.е. независимости s_1 и s_2.

Значение этой функции-меры больше для предложений, исключающих большее количество возможностей. Пример: из s_1 — "a > 3" и s_2 — "a = *7*' следует, что s_2 => s_1 или *inf*(s_2) >= inf(s_1); ясно, что s_2 исключает больше возможностей, чем s_1.

Прагматическая мера информации

Эта мера определяет полезность информации (ценность) для достижения пользователем поставленной цели. Эта мера также величина относительная, обусловленная особенностями использования этой информации в той или иной системе.

Качество информации

Возможность и эффективность использования информации обусловливаются такими основными ее потребительскими *показателями качества,* как репрезентативность, содержательность, достаточность, доступность, актуальность, своевременность, точность, достоверность, устойчивость.

Репрезентативность информации связана с правильностью ее отбора и формирования в целях адекватного отражения свойств объекта. Важнейшее значение здесь имеют:

- правильность концепции, на базе которой сформулировано исходное понятие;
- обоснованность отбора существенных признаков и связей отображаемого явления.

Нарушение репрезентативности информации приводит нередко к существенным ее погрешностям.

Содержательность информации отражает семантическую емкость, равную отношению количества семантической информации в сообщении к объему обрабатываемых данных, т.е.

$$C = \frac{I_C}{V_Д}$$

С увеличением содержательности информации растет семантическая пропускная способность информационной системы, так как для получения одних и тех же сведений требуемся преобразовать меньший объем данных.

Наряду с коэффициентом содержательности С, отражающим семантический аспект, можно использовать и коэффициент информативности, характеризующийся отношением количества синтаксической информации (по Шеннону) к объему данных

$$Y = \frac{I}{V_Д}$$

Достаточность (полнота) информации означает, что она содержит минимальный, но достаточный для принятия правильного решения состав (набор показателей).

Понятие полноты информации связано с ее смысловым содержанием (семантикой) и прагматикой. Как неполная, т.е. недостаточная для принятия правильного решения, так и избыточная информация снижает эффективность принимаемых пользователем решений.

Доступность информации восприятию пользователя обеспечивается выполнением соответствующих процедур ее получения и преобразования. Например, в информационной системе информация преобразовывается к доступной и удобной для восприятия пользователя форме. Это достигается, в частности, и путем согласования ее семантической формы с тезаурусом пользователя.

Актуальность информации определяется степенью сохранения ценности информации для управления в момент ее использования и зависит от динамики изменения ее характеристик и от интервала времени, прошедшего с момента возникновения данной информации**.**

Своевременность информации означает ее поступление не позже заранее назначенного момента времени, согласованного со временем решения поставленной задачи.

Точность информации определяется степенью близости получаемой информации к реальному состоянию объекта, процесса, явления и т.п.

Достоверность информации определяется ее свойством отражать реально существующие объекты с необходимой точностью. Измеряется достоверность информации доверительной вероятностью необходимой точности, т.е. вероятностью того, что отображаемое информацией значение параметра отличается от истинного значения этого параметра в пределах необходимой точности.

Устойчивость информации отражает ее способность реагировать на изменения исходных данных без нарушения необходимой точности.

Источники

1. Лидовский, Теория информации, 2004. Разделы 5,7,8,9.

2. Конспект лекций УГТУ по информатике стр. 10-17