Информационная биология

- Тема 4

- Количественное оценивание информации

Подход к оцениванию информации

- В общей теории информации формальный аппарат для оценки количества информации выделяет, различает три аспекта:
- А) статистический;
- Б) семантический;
- В) прагматический.

Статистический аспект

- Статистический аспект информации был разработан применительно к целям и задачам теории связи К. Шенноном. Теория связи оперирует знаками и абстрагируется от семантического и прагматического аспектов.
- Статистическая информация не делает различий между важной, новой информацией, её ценностью, полезностью для того, кто её получает. Такой подход делает количество информации объективной оценкой, но она становится безликой.
- *Количество информации определяется на основе понятий теории вероятностей, путём оценивания вероятности появления того или иного сигнала, знака, буквы алфавита*

Информационная энтропия

- Основным, базовым понятием при количественном оценивании информации является **энтропия (информационная).**
- Энтропия (физическая) – мера рассеяния энергии в тепло в замкнутой термодинамической системе (Клаузиус, Больцман, 1852).
  - **S = k ln W [э.е.],**

  где S – т.д. энтропия, W – число состояний системы.
- **1 [э.е.] = 1 кал/град = 4.2 Дж/К**

Информационная энтропия

- **Энтропия – мера вероятности информационных систем** (Л. Сциллард, К. Шеннон, 1929)
- **Энтропия – мера дезорганизации систем любой природы** (Шрёдингер, 1944; Л.Бриллюэн, Н. Винер, ?)
- Информация и энтропия могут быть связаны соотношением **H + I = const**.
- H-мера беспорядка, I-мера упорядоченности.
- **1 э.е. = 2.3*10 (-23) бит**

Информационная ёмкость

- Количественная мера информации должна отвечать требованию аддитивности.
- В 1928 г. Хартли предложил оценивать информационное содержание систем как ***логарифм числа возможных состояний*** системы, назвав это «информационная ёмкость .
- **C = log N = - log 1/N = - log P = H**
- Информационная ёмкость нашей аудитории…….

Информационная ёмкость

- Если информационная система может находиться в **N** возможных состояниях и все они взаимно независимы, то информационная ёмкость такой системы равна **C = log N.**
- Две таких системы будут иметь N в квадрате состояний , т.к. каждому состоянию 1-й системы будут соответствовать N возможных состояний 2-й. Информационная ёмкость двух систем будет равна

- C1 + C2 = log (N 2) = 2 log N = 2C

Информационная ёмкость

- Т.о., информационная ёмкость проявляет свойство **аддитивности** и, в общем, ёмкость k систем будет в k раз больше ёмкости одной системы.
- **Информационная ёмкость по-другому называется *мера Хартли*.** При расчёте может использоваться логарифм с любым основанием (десятичный, натуральный), но это привносит некую неопределённость при использовании в расчётах.

Информация по-Шеннону

- К. Шеннон предложил при расчёте энтропии использовать **логарифм с основанием 2**. Он исходил из необходимости решения задачи о количестве информации в каналах связи при условии: есть сигнал или его нет (два варианта состояния, двоичное кодирование).
- При равновероятных событиях «есть-нет» количество энтропии как меры неопределённости (или информации, нужной для устранения этой неопределённости), будет *равна числу необходимых двоичных выборов*. Математически это равно *двоичному логарифму числа состояний.*

Информация по-Шеннону

- $H = \log_2 N = -\log_2 P$,

H – количество энтропии в битах [бит],

N - число равновероятных состояний системы,

P – вероятность нахождения системы в некотором состоянии.

Для кубика с разным числом точек на каждой грани вероятность нахождения в каждом из состояний равновелика (1/6).

В реальности состояния и свойства систем характеризуются разными вероятностями (что-то более, что-то менее вероятно).

Поэтому для определения энтропии не равновероятных событий используется более сложная формула, учитывающая сумму вероятностей всех событий.

Информация по-Шеннону

- $H = - \Sigma p(i) \log_2 p(i)$
- Размерность энтропии – [бит /символ].
- Смысл – сколько надо сделать двоичных выборов (вопросов), чтобы снять, устранить неопределённость величиной H.
- Для определения количества полной информации в тексте сообщения необходимо
- $I = - N \Sigma p(i) \log_2 p(i)$, N – число символов в сообщении.
- Размерность [бит];
- 1 бит- количество информации, необходимой для передачи или хранения двоичного

Информация по-Шеннону

- Энтропия, рассчитанная для равновероятных событий, может считаться как априорная H (апр), в то время как рассчитанная для не равновероятных событий считается апостериорной (после опыта) – действительно, вероятности-то определены опытным путём) H (апост).
- Поэтому иногда информация определяется как устранённая неопределённость по формуле **I = H (апр) – H (апост)**

Пример со студентами

- На занятии 32 студента. Один из них поощрён и преподаватель должен определить его, не вызывая по фамилии .
- **Сколько вопросов с ответами «да-нет» должен задать преподаватель, чтобы идентифицировать студента?**
- Эта процедура идентична оцениванию информации в битах.

Алгоритм

- Список студентов, составленный по алфавиту без учёта пола, делится пополам и уточняется, есть ли студент в первой половине. Круг поиска суживается вдвое (с 32 до 16). Действуя аналогично число претендентов сужается и 5-м вопросом преподаватель узнаёт необходимое.
- $H = \log_2 32 = 5$ [бит/символ «студент»]

Алгоритм

- Общее правило:
- **если есть N элементов и один из них X как-то должен быть обнаружен, то для этого необходимо иметь информацию, достаточную чтобы устранить неопределённость H.**
- **$H = \log_2 N$.**
- Эту величину можно считать мерой Хартли, оцененной в битах.

Английский алфавит как объект количественного оценивания

Частота встречаемости букв английского алфавита [26]

Символ	Вероятность	Символ	Вероятность
Интервал между словами	0,2	L	0,029
E	0,105	C	0,023
T	0,072	FU	0,0225
O	0,0654	M	0,021
A	0,063	P	0,0175
N	0,059	YW	0,012
I	0,055	G	0,011
R	0,054	B	0,0105
S	0,052	V	0,008
H	0,047	K	0,003
D	0,035	X	0,002
		JQZ	0,001

Резюме

- Какое количество энтропии (информации) содержится в сообщении на основе букв английского алфавита?
- **Если все буквы передаются с одинаковой вероятностью, то p =1/27;**
- **$H = -\log_2 1/27 = 4.76$ бит/символ(букву).**
- Это аналогично тому, что необходимо 5 ячеек памяти для 0 и 1. Или же **необходимо задать 5 вопросов «да-нет»,** чтобы определить любую искомую букву алфавита.

Резюме

- Поскольку реальная вероятность использования разных букв разная, то с учётом этого обстоятельства H = 4.03 бит/символ. Т.е. число двоичных ответов, необходимых для идентификации буквы уменьшилось.
- H уменьшается ещё больше, если учесть наличие дифтонгов или трифтонгов (th, tch), кода вероятность появления определённых букв после t возрастает. С учётом этого энтропия понижается до H = 3.35 бит/символ для 2-х букв и H = 3.1 бит/символ для 3-х.

Резюме

- С учётом всех особенностей английского языка **H = 1.5 бит/символ.**
- Пример с английским алфавитом иллюстрирует два важных положения статистической информации: а) когда все вероятности знаковых событий равны, количество энтропии H максимально; б) если вероятность данного сообщения (знака) связана с вероятностью появления другого сообщения (знака), величина энтропии H уменьшается

Основные понятия статистической теории информации

- **Информационная ёмкость** сообщения – характеризует источник сообщения;
- **Избыточность символов** – характеризует источник сообщения;
- **Пропускная способность** канала связи – характеризует канал связи;
- **Надёжность, помехоустойчивость** – характеризует всю информационную систему в целом.

Информационная ёмкость

- Если текст содержит N символов, то информационная ёмкость рассчитывается по формуле Шеннона: H = **I = -Σ P(i)log$_2$ P(i);**
- Или же: Информационная ёмкость - это количество информации в битах, содержащейся в оцениваемом сообщении
- **I = H(апр) – H(апост)**

Избыточность информации, символов

- Можно писать текст сокращёнными словами (лекции), но смысл фраз оказывается вполне понятен.
- Полностью записанный текст содержит больше символов, чем требуется для однозначного понимания содержания.
- **Наличие чрезмерного количества знаков для написания сообщения называется избыточностью и может измеряться в битах**

Избыточность

- В английском языке H = 1.5 бит/символ, в то время как H(ср.) = 4.7 бит/символ. Получается, что 3.2 бит/символ лишние
- Для чего необходима избыточность? *Избыточность знаков, сообщений необходима как условие, препятствующее появлению искажения, ошибок.*
- При отсутствии избыточности любой сбой в системе связи приводит к возможности появления не обнаруживаемых и не исправляемых ошибок в принятой информации.
- Такая информация – зашумлённая,
- искажённая.

Расчёт избыточности

- Информационная избыточность может быть рассчитана через относительную энтропию h
- **h = H(эмп.) / H(макс.). Для англ. алф.**
- **h = 1.5/4.7 = 0.32.**

- **Величина относительной энтропии h используется для оценки избыточности**
- **D = 1 – h = 1 - 0.32 = 0.68.**

- Избыточность – безразмерная величина

Смысл избыточности

- Избыточность употребляется в том смысле, что часть информации не является необходимой для передачи и понимания смысла сообщения.
- Избыточность по Шеннону - это техническое понятие в теории информации для количественной технической оценки избыточности.
- Но, вычисленное значение избыточности не всегда можно соотнести с конкретным содержанием или пониманием чего-либо.

Избыточность и генетический код

- Избыточность конкретной молекулы ДНК необходимо оценивать с учётом ограничений, связанных с частотой встречаемости определённых соседних нуклеотидов. Если есть данные по всем 4 основаниям ДНК, то можно корректно оценить избыточность.
- Л. Гетлин вычислила избыточность ДНК разного происхождения и обнаружила, что она очень низка, в пределах 0 – 11%.
- Однако у некоторых ДНК избыточность велика. Сателитная ДНК краба имеет избыточность 83%. ДНК некоего вируса имеет следующий состав: А = 87%, Т = 10,5%, Ц = 1,4%, Г = 0,4%. Это приводит к очень высокой избыточности генетической информации.

Избыточность и генетический код

- Если в молекуле ДНК пропущен или изменен один нуклеотид, то биологические последствия в большинстве случаев могут быть очень серьёзные.
- Рассмотрим пример с последовательностями
- -А-А-Г-Г-Г-**У**-Ц-Ц-А-У-Ц-А-Ц-У-У-А-А-
- -А-А-Г-Г-**У-**Ц-Ц-А-У-Ц-А-Ц-У-У-А-А-
- Такая мутация происходит в ДНК фага Т4.
- В результате последовательность АК в молекуле кодируемого белка лизоцима
- -Лиз-Сер-Про-Сер-Лей—Асп-Ала- меняется
- -Лиз-Вал-Гис-Лей—Мет-Ала-
- Образуется белок с другими свойствами.

Белки и избыточность

- -Вал-Гис-Лей-Тре-Про-**Глу**-Глу- норма в г-г
- -Вал-Гис-Лей-Тре-Про-Вал-Глу- замена одной АК в гемоглобине приводит к серповидно-клеточной анемии.
- Поэтому понятие генетической или белковой избыточности имеет другой характер, чем в технической теории информации. Пропуск одного «слова» полностью обесценивает сообщение, которое нельзя исправить.
- Поэтому в процессе передачи генетической информации есть системы исправления ошибок на уровне тРНК.
- Аминоацил-АМФ и Аминоацил-тРНК.

Пропускная способность

- Пропускная способность связана со скоростью передачи информации.
- Пропускная способность среды (канала) – **максимальное количество единиц информации (бит),** которые данная среда (канал) может ***безошибочно*** **пропустить через себя в единицу времени.**
- П.С. канала связи – максимальная скорость *безошибочной* передачи сигнала (информации) в данной среде, измеряемая в бит/ сек.
- В общем случае п.с. канала определяется:
- **C = 1/T max I (X – Y) [бит/сек];**
- **$C = B \log_2 (1 + S/N)$ [бит/сек];**
- **B – полоса сигнала, Гц; S – средняя мощность**

Помехоустойчивость, надёжность

- Помехоустойчивость, надёжность информационных систем – способность *безошибочно* генерировать, передавать, запоминать и воспроизводить информацию.
- Мера надёжности передачи сообщения выражается следующим образом:

$$S = \log_2 1/ P(0),$$

где P(0) – вероятность ошибочной передачи сигнала

Надёжность ж.с.

- Живые системы характеризуются высокой надёжностью функционирования. Формально надёжность живых систем определяется следующим:
- **S = 1/P(0),** где P(0) – вероятность нарушения функции системы.
- При P(0) – минимум, S – максимальна.

Надёжность ж.с.

- Живые системы – высоко надёжны.
- Надёжность ж.с. во многом определяется дублированием элементов или функций. Т.е. одновременно процесс выполняется параллельными элементами и число их избыточно.
- При повреждении или необратимой утрате некоторого количества клеток, органов, объектов цель, результат их функционирования не пострадает.
- Примеры: избыточное число нуклеотидов, нервных волокон в нерве, избыточные кладки яиц, семян и т.д.

Примеры использования статистической информации

- Одним из первых, оценивших потенциальные возможности теории информации, был Г. Кастлер, который в 1955 г. издал книгу о биологических приложениях этой теории. В частности, Кастлер подсчитал, что ДНК млекопитающих обладает информационной ёмкостью 2 * 10(10) бит
- Это эквивалентно информации 100 комплектов Британской энциклопедии.

Примеры

Нейроны	H	F
Helix	1.1-2-0 бит	1.2 ПД/с
Речной рак	2 – 3.7 бит	2.2 ПД/с
Лягушка	1.4-3.7 бит	4.1 ПД/с
Крыса	2.9-4 бит	5.0 ПД/с
Кролик	3 -4.7 бит	5.8 ПД/с

- Имеет место постепенное нарастание фонового импульсного потока по мере эволюционного совершенствования нервных структур, информационные возможности увеличиваются.

Ограниченность использования статистической информации

- «…Большинство работ с применением теории информации в биологии тривиальны – известные факты и положения переводятся на другой язык» - Л. А. Блюменфельд
- Действительно, шенноновская теория информации, рассматривает вопросы только о её количестве.
- В основе теории не сколько о количестве информации, сколько об информационной ёмкости «тары» - совокупности знаков, символов, предназначенных для хранения или передачи информации.
- Содержание, смысл, ценность информации при этом не учитываются…..
- Но, оценивание проводится в одном масштабе, что позволяет адекватно сравнивать, сопоставлять разные информационные объекты и процессы.