FAQ. Теорема Шеннона

Q. Теорема Шеннона

Согласно теореме Шеннона "Элемент Si, вероятность появления которого P(Si), выгоднее всего представить - Log2(P(Si)) битами"

То есть, это минимальное значение.

В некоторых случаях при кодировании методом Хаффмана получаются коды, короче рассчитанных по приведенной выше формуле. Почему так получается?

A. По этой формуле получается средняя длина кода. Коды Хаффмана (неравномерное кодирование) начинаются с однобитных - и дальше вверх. При этом средняя длина кода, обычно, больше теоретической.

IMHO, использование в ТИ аппарата теории вероятностей (влияние Н.Винера?) только препятствует пониманию.

Я бы начал с деревьев (выбора). При двоичном выборе, расщепление пространства состояний пополам отвечает продвижению на один уровень (один шаг) по дереву выбора.

Например, имеется упорядоченная последовательность из восьми чисел: 1..8, в которой загадано одно число.

Сколько вопросов, ответ на которые Да / Нет, нужно задать для отгадывания числа? Методом деления пополам, разбиваем пространство состояний 1..8 на блоки по 4, по 2, по 1 - число угадано.

Иными словами, за 3 шага по дереву приходим из корня в терминальный узел (лист).

Это дерево идеально сбалансировано (дерево Хартли) и длина пути из корня ко всем узлам одинакова:

H = LOG2(8) = 3

Но, Шеннон избавляется от абсолютных значений (кратностей) и переходит к относительным (частотам), а, кроме того, нормирует размер пространства состояний, принимая его за 1 (100%). Таким образом, появляется "вероятность".

Для сбалансированного дерева, вес каждого узла будет 1/8 полного веса, и это и есть то самое Pi

Иными словами, от "прямых" абсолютных единиц подсчета мы перешли к "обратным" относительным.

Но само дерево от этого, естественно, не изменилось. Его высота, по-прежнему, три, но выражается через обратные величины:

H = -LOG2(1/8) = 3

Пока все это мало отличается от того, что было предложено Хартли, не считая неудобной формы записи.

И вот здесь, Шеннон вносит революционное предложение: вводит в рассмотрение несбалансированные деревья.

При этом, как всякий математик, старается свести задачу к уже известной - путем введения поправочных коэффициентов.

Если деревья Хартли были "прямыми" - в том смысле, что длина (вес) всех дуг между узлами была одинакова (1), то Шеннон строит "косое" дерево, с дугами неравной длины (неравными весами).

Для несбалансированного дерева, Шеннон вводит компенсирующие весовые коэффициенты.

Если представить теперь дерево как систему трубок (сосудов), где втекающий в корень поток жидкости разветвляется по более мелким трубкам, то эти коэффициенты выбираются из условий неразрывности потока: сколько в корне втекло, столько в сумме из всех листьев вытекло.

Запись этого условия (Закона сохранения) для потока вероятности и есть знаменитая формула Шеннона.

***

Понятно, что для "косого" дерева, длина пути от корня к листу может и не быть целым числом.

Но, когда используется целочисленное кодирование (например, коды Хаффмана), длина любого кода выражается целым числом бит. Иными словами, в общем случае, дерево Хаффмана не будет совпадать с деревом Шеннона, но будет близко к нему.

Как удалось доказать Хаффману (в то время, аспиранту Фано), не существует других целочисленных кодов (деревьев), которые отличались бы от дерева Шеннона меньше, чем дерево Хаффмана.

Таким образом, код Хаффмана является кодом с минимальной избыточностью (наилучшим дискретным приближением к непрерывному представлению Шеннона).

При этом, длина конкретного целочисленного кода, естественно, может быть как меньше, так и больше парного ему точного значения для непрерывного кода. Это НЕ ошибка формулы Шеннона - это ошибка аппроксимации точных непрерывных объектов кусочно-линейной моделью (a la вписанным/описанным в круг многоугольником).