

Нормализация дискретных данных

- **Индикатор** — длина закодированной строки равна $n-1$, где n — количество уникальных значений. Каждому уникальному значению поля, кроме опорной категории, выделяется своя колонка, в которую записывается значение 1, а все остальные значения строки равны 0; опорная категория кодируется строкой из нулей.
- **Индикатор (без опорной категории)** — длина закодированной строки равна n . Опорное значение не используется.
- **Отклонение** — то же, что и "Индикатор", но опорная категория кодируется строкой из -1.
- **Простая** — то же, что и "Индикатор", но вместо 0 используется значение $-1/n$, а вместо 1 используется значение $1-1/n$.
- **Разность** — уникальные значения упорядочиваются, после чего каждое уникальное значение сравнивается со следующим по таблице вида:

Уникальные значения	Контраст 1	Контраст 2	Контраст 3
	U1 vs. U2	U2 vs. U3	U3 vs. U4
U1	$(n-1)/n$	$(n-2)/n$	$(n-3)/n$
U2	$-1/n$	$(n-2)/n$	$(n-3)/n$
U3	$-1/n$	$(n-2)/n$	$(n-3)/n$
U4	$-1/n$	$-2/n$	$-3/n$

- **Обратная разность** — уникальные значения упорядочиваются, после чего каждое уникальное значение сравнивается с предыдущим по таблице вида:

Уникальные значения	Контраст 1	Контраст 2	Контраст 3
	U1 vs. U2	U2 vs. U3	U3 vs. U4
U1	$-(n-1)/n$	$-(n-2)/n$	$-(n-3)/n$
U2	$1/n$	$-(n-2)/n$	$-(n-3)/n$
U3	$1/n$	$2/n$	$-(n-3)/n$
U4	$1/n$	$2/n$	$3/n$

- **Гельмерт** — уникальные значения упорядочиваются, после чего каждое уникальное значение сравнивается со средним из последующих значений по таблице вида:

Уникальные значения	Контраст 1	Контраст 2	Контраст 3
	U1 vs. U2, U3 и U4	U2 vs. U3 и U4	U3 vs. U4
U1	$(n-1)/n$	0	0
U2	$-1/n$	$(n-2)/n$	0
U3	$-1/n$	$-1/(n-1)$	$(n-3)/n$
U4	$-1/n$	$-1/(n-1)$	$-1/(n-2)$

- **Обратный Гельмерт** — уникальные значения упорядочиваются, после чего каждое уникальное значение сравнивается со средним из предыдущих значений по таблице вида:

Уникальные значения	Контраст 1	Контраст 2	Контраст 3
	U2 vs. U1	U3 vs. U1 и U2	U4 vs. U1, U2 и U3
U1	$-1/(n-2)$	$-1/(n-1)$	$-1/n$
U2	$(n-2)/n$	$-1/(n-1)$	$-1/n$
U3	0	$(n-2)/n$	$-1/n$
U4	0	0	$(n-1)/n$

- **Индекс уникального значения** — каждое уникальное значение заменяется на индекс уникального значения.

Опорное значение

В некоторых нормализаторах дискретных данных используется опорное значение. Выбор опорного значения происходит из таблицы уникальных значений. Существует несколько вариантов задать опорное значение:

- Первое в списке;
- Последнее в списке;
- Более редкое;
- Более частое;
- Задано явно.