Меры центральной тенденции

Антон Антонов, Андрей Ковалевский, Лишуди Александр, Киракосян Александр

10 декабря 2021 г.

Аннотация

В данном произведении будут рассмотрены 4 дополнительных темы, относящихся к мерам центральных тенденций: различные средние, не освященные в оригинальном видео, среднее значение, среднее, медиана и мода сгруппированных данных, медианный фильтр, фильтр скользящего среднего.

1 Introduction

Предположим, что Ваш знакомый постоянно говорит Вам фразу 'Купи слона', и Вам это надоело. Чтобы прекратить это, Вы решили и вправду купить слона. Однако, для начала надо узнать, насколько это дорого. Допустим, поиск в интернете возможностей покупки слона в Вашем городе дал следующий результат: средняя цена слона — 420 тысяч рублей. Кажется, что это очень дорого, но не повод отчаиваться — ведь как утверждают знающие люди, лучше всего смотреть не на среднее значение, а на медианное, а в некоторых случаях — на моду. Всё это называется мерами центральной тенденции. Попробуем разобраться, что же это такое.

Меры центральной тенденции — это величины, позволяющие одним числом передать главную информацию о выборке. Всего их три: среднее, медиана и мода.

1. Среднее значение выборки — это её сумма, делённая на её размер. Оно довольно осмысленно описывает выборку, если она распределена приблизительно нормально. Но у него есть один существенный минус — чувствительность к выбросам.

2. Медиана — это центральный элемент отсортированной по возрастанию выборки. Например, если у Вас есть два кота, у Джилиана есть один кот, а у Эрика три кота, то медиана здесь — это два кота. Если добавить в выборку Уилла, у которого 14 котов, то посередине оказывается уже два значения - 2 и 3. Обычно в таких случаях говорят, что медиана равна среднему между ними, то есть 2.5 котам. Медиана хороша в тех случаях, когда в данных имеются значительные выбросы — отличным примером тагоко случая будет ситуация, когда есть компания из 10 человек с обычным доходом и Илона Маска, и надо оценить, каковы будут траты этой компании на еду в ресторане.

3. Мода — это самое частое значение в выборке. Мода бывает очень полезна в тех случаях, когда выборка состоит не из чисел — например, если есть выборка из тех напитков, которые люди пьют по утрам, то для неё нельзя найти ни среднее, ни медиану (ведь мы не можем ни складывать, не упорядочивать напитки), но можем найти моду — напиток, который чаще всего пьют по утрам (кофе).

1.1 Median Filter

Где, помимо решения статистических задач, применяются меры центральной тенденции? Например, в обработке изображений. Пожалуй, самым ярким примером здесь будет медианный фильтр. Допустим, требуется отсканировать фотографию. Вы пришли к другу, у которого есть сканер, положили на стол фотографию, и позвали друга, чтобы он её отсканировал. Друг предложил для начала пообедать. В процессе обеда вы просыпали на фотографию соль, а когда потянулись за упавшей солонкой, задели перечницу, и перец тоже просыпался на фото. Пока Вы пытались отыскать упавшую перечницу, Ваш не очень внимательный друг взял фотографию и положил

её в сканер прямо с солью и перцем, попавшими на неё, отсканировал и вернул Вам. Пока друг шел от сканера к Вам, влетевший сквозь открытое окно порыв ветра сдул с фото и соль, и перец, так что Вы не заподозрили, что что-то пошло не так. Когда после длительной поездки через весь город Вы наконец вернулись домой и стали смотреть входящие письма на электронной почте, вы обнаружили, что скан фотографии получился плохим из-за соли и перца, находившихся на ней в процессе сканирования, и теперь хотите убрать эти загрязнения с фотографии. Такая ситуация кажется невероятной, однако на деле точно такие же загрязнения (называемые шумом типа соль и перец) довольно часто генерируются светочувствительными матрицами фотоаппаратов. Чтобы убрать этот шум, надо применить фильтр. Опишем работу фильтра для серой фотографии, а для цветной будем применять фильтр к каждому из трёх цветовых каналов отдельно. Наиболее очевидным будет фильтр арифметических средних: для каждого квадрата (например, размера 3×3) из пикселей значение его центрального пикселя заменяется на среднее значений пикселей в квадрате. Те, кто изучали нейронные сети, знают, что эта операция является частным случаем свёртки, в котором веса при всех пикселях в квадрате 3×3 (его ещё называют ядром) одинаковы. Однако, такой фильтр будет не очень полезен, что видно на примере ниже.

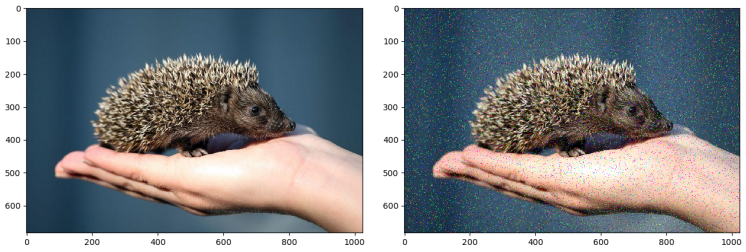
Возьмём вот такую фотографию: 1a и применим к ней шум типа "соль и перец". Получится следующее изображение (так как мы применили шум к каждому цветовому каналу отдельно, получились отчётливо видны яркие синие, зелёные и красные точки): 1b Теперь применим фильтр арифметических средних: 1c Как можно заметить, картинка всё ещё зашумлена, хоть и не настолько сильно. Это весьма логично — если писели, скажем, зелёного канала, в квадрате 3×3 имеют значения 9, 9, 10, 10, 12, 13, 14, 15, 255, то среднее арифмитическое от них будет равно 43 — значение куда выше, чем большинство из имеющихся, так что в тёмных облястях фото (в данном случае, на фоне) шум не будет ликвидирован до конца. И наоборот, в очень светлых областях (рука, дежращая ежа) остаются точки темнее, чем сама рука.
Теперь применим медианный фильтр — это фильтр, которые как и фильтр арифметических средних заменяет каждый пиксель значением, посчитанным в квадрате 3×3 вокруг него, но теперь для этого квадрата будет считать медиану в нём. Вот что получилось: 1d Как видно, медианный фильтр не оставил ни следа от шума. Почему так произошло? Посчитаем медиану для того же квадрата, для которого мы считали среднее арифмитическое — медиана значений 9, 9, 10, 10, 12, 13, 14, 15, 255 равна 12 — и это число куда больше похоже на остальные числа из этого квадрата, нежели 255. Так происходит потому что 255 — это выброс, а медиана устойчива к выбросам. Вот так медиана помогла нам восстановить фотографию ежа, хотя это и казалось невозможным.

1.2 Moving average

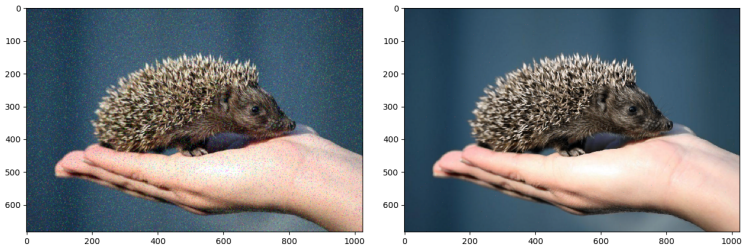
До этого были рассмотрены меры центральной тенденции для разных выборок. Но очень часто приходится работать с временными рядами, в них тоже хотелось бы иметь какой-то аналог центральной тенденции во времени. Для данного случая как раз был придуман фильтр скользящего среднего, более известный на английском как moving average или running mean.
Можно добавить что данный фильтр может помочь нам убрать мешающие увидеть тенденцию небольшие колебания.
Суть же данного фильтра состоит в том, что он берет подряд значения временного ряда и сглаживает их, то есть усредняет, по настоящей и предыдущим n - 1 точкам:

$$y_i = \frac{1}{n} \sum_{k=0}^{n-1} y_{i-k} \tag{1}$$

Данный фильтр параметризуется всего одним числом - n, число предыдущих точек, с помощью которых мы рассчитываем новое значение. Подбирать надо аккуратно, так как слишком большой размер - сильно сгладит, а маленький может наоборот не до конца сгладить.
Стоит продемонстрировать работу данного фильтра:
Возьмем данные, сгенерированные случайным блужданием (зеленый), предположим, что это запасы яблок ежика из предыдущего задания каждый месяц, но так как ежик не посещал школу, он неправильно их считает (оранжевый), то прибавит лишка, то недосчитает. Однажды ежика посетил аудитор леса, чтобы узнать, как ему живется последние 10 лет. Ежик поднял свою бухгалтерию, которая невероятно поразила аудитора. Но работу все равно надо было закончить,



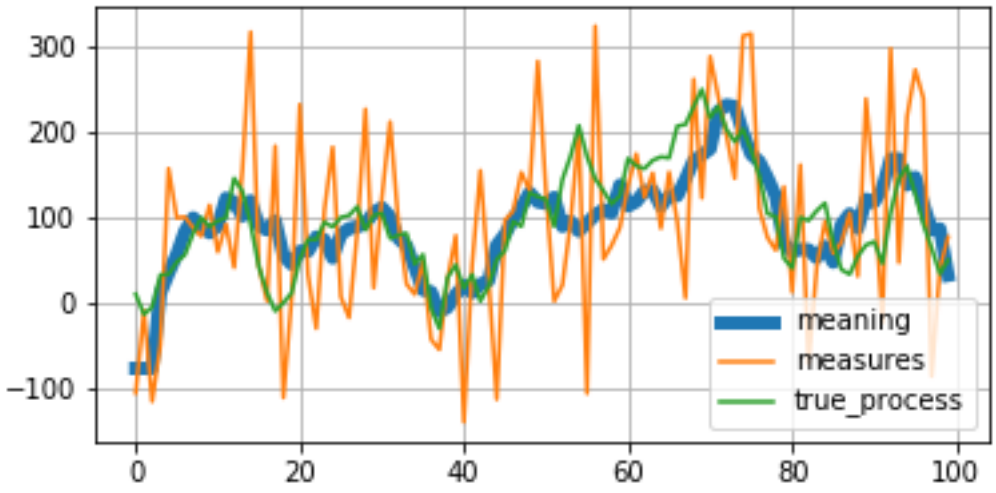
(a) Ежик (b) Шум типа "соль и перец"



(c) фильтр арифметических средних(d) медианный фильтр — возвращение ежа

Рис. 1: Ёжики без тумана

проверяющий понял, что ежик примерно одинаково часто прибавлял и убавлял яблоки, и с помощью скользящего среднего смог получить примерную картину (синий), которая достаточно неплохо аппроксимировала правдивую.



1.3 Central tendency of grouped data

Иногда необходимо посчитать центральные тенденции для сгурппированных по диапозонам данных, т.е. когда известно сколько раз значение попало в каждый из диапозонов.

Практически диапозоны могут возникать из-за изначальной смысловой группировки данных, например при группировке результатов тестов учащихся по баллам для выставления оценки. При такой поставновке нет информации о том сколько конкретно баллов набрал каждый ученик, а

Балл	Количество
40 - 49	3
50 - 59	5
60 - 69	6
70 - 79	9
80 - 89	8
90 - 100	7

значит точное значение центральных тенденций посчитать невозможно. Тем не менее, возможно посчитать некоторые приближения.

Пример диапозонных данных приведен в таблице

Покажем, как посчитать среднее, медиану и моду таких данных. Общим подходом здесь является предположение, что задача аналогична поиску центральных тенденций для обычного числового ряда, образованного при помощи замены диапозона на конкретное среднее значение из него.

Например вместо первой строки таблицы в исходный числовой ряд будет три раза добавлено число 44.5, являющееся серединой интервала. Аналогично для последней строки будет 7 раз добавлено число 95.

Перейдем к упрощенному описанию подсчета:

- Для подсчета моды аналогично числовому ряду достаточно взять самый частый диапозон, который и будет являться ответом.
- Для подсчета медианы необходимо понять в каком из интервалов будет лежать центральное значение. По итогу это аналогично поиску медианы описанного ранее ряда, а затем взятие соответсвующего интервала.
- Для подсчета моды необходимо сложить средние значения домноженные на частоты и разделить на сумму всех частот. Формульно это выглядит как $\frac{\sum f_i \cdot m_i}{\sum f_i}$. При детальном рассмотрении, можно заметить, что это в точности совпадает со средним значением числового ряда, описанного ранее.

1.4 Geometric Mean and Harmonic Mean

Для некоторых видов данных использование среднего арифметического приводит к получению неверных выводов. Например, если вы работаете с финансовыми данными, то там часто встречаются данные о росте какой либо величины в процентном соотношении. Если мы захотим посчитать рост такой величины за какой либо период времени, то нам придется перемножать проценты, а не складывать. Поэтому использование арифметического среднего в данном случае даст неправильный результат. Нужно использовать среднее геометрическое: $\prod_{x=1}^{n} x_i$.

Давайте рассмотрим пример: вы инвестируете и записываете прибыльность вашего портфеля, первые три года ваша прибыль составляет 10%, а на четвертый вы покупаете биткоин и получаете 200% прибыли. Среднее арифметическое составит $\frac{1.1 + 1.1 + 1.1 + 3}{4} = 1.575$, а среднее геометрическое $\sqrt[4]{1.1 + 1.1 + 1.1 + 3} = 1.41$ разница в 16% довольно велика.

Иногда может пригодится и другое среднее - гармоническое. Допустим вы занимаетесь бегом и хотите проанализировать ваши результаты. Вы знаете что на тренировке вы пробежали 2 круга со скоростью 10км/ч и 2 круга со скоростью 20км/ч. Если вы возьмете среднее арифметическое, то получите 15км/ч. Однако, получится что вы усредняете скорость по дистанции, а не по времени. На деле, на два круга со скоростью 20 км/ч вы потратили в два раза меньше времени, чем на два круга со скоростью 10км/ч. Здесь нужно воспользоваться средним гармоническим: $\frac{n}{\sum_{x=1}^{n} (\frac{1}{x_i})}$

тогда мы получим корректный ответ 13.(3).

1.5 References

Median filter, Moving average, Central tendency of grouped data, Geometric Mean and Harmonic Mean, help library