ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Анализ и преобразование данных

Высшая Школа Цифровой Культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Описательная статистика

Описательная статистика – это первичная систематизация данных, полученных из различных источников. Описательная статистика – раздел статистической науки, в рамках которого изучаются методы систематизации, описания и представления основных свойств данных. Важно отметить, что описательная статистика работает только с конкретной выборкой и не стремится охарактеризовать какие-либо свойства генеральной совокупности, поэтому определения и понятия, используемые в описательной статистике, иногда отличаются от определений, даваемых в индуктивной статистике. Описательная статистика включает в себя: сбор данных, их категоризацию, обобщение и представление. Описательная статистика активно используется на этапе разведочного анализа данных, а в некоторых случаях вообще оказывается достаточной для полного анализа данных. Рассмотрим основные виды описательных статистик и их практическое применение.

Мерой центральной тенденции в описательной статистике принято называть число, описывающее все значения выбранной переменной из набора данных. Мера центральной тенденции позволяет описать типичную выраженность признака рассматриваемого набора данных. Чаще всего выделяют следующие характеристики центральной тенденции:

- среднее значение;
- мода;
- медиана.

Среднее значение переменной определяют как сумму всех значений переменной, деленную на количество значений. Обозначается $\overline{x}$ или Mean.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Среднее вычисляется только в числовых шкалах и в дихотомических данных с 0 и 1. Для каждого набора данных имеется только одно среднее. Рассмотрим пример вычисления среднего для отметок студента. У студента в процессе его обучения в университете были получены следующие отметки:

$$5, 4, 2, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5$$

Среднее будет вычислено как сумма всех значений деленная на количество, т.е.:

$$\mathsf{Mean} = \frac{58}{14} \approx 4.14$$

Как уже было сказано, среднее может также вычисляться и для дихотомических данных. Если два значения переменной представляются 0 и 1, то среднее для таких данных указывает долю единиц в выборке. Например, для следующих данных:

$$1, 0, 0, 0, 1, 1, 1, 0, 0, 0,$$

40% значений выборки принимают значение, равное единице:

$$\mathsf{Mean} = \frac{4}{10} = 0.4$$

Мода — это значение переменной, которое встречается чаще других. Обозначается Mo. Мода может быть определена на данных любой шкалы. Моде может соответствовать несколько значений. В этом случае говорят про мультимодальное распределение значений переменной. Если частота всех значений признака в наборе данных одинакова, то говорят, что либо мода отсутствует, либо все значения являются модой. На рисунке вы можете видеть пример вычисления моды для отметок студента. Наиболее частая отметка – 5.

Отметка	Частота
5	7
4	3
3	3
2	1

Mo = 5

Еще одна характеристика центральной тенденции – медиана. **Медиана** – это такое значение признака, что ровно половина значений признака не больше него, а другая половина не меньше него. Чтобы определить медиану, удобно ввести понятие вариационного ряда.

Медиана

Обозначение – Me

Чтобы определить медиану, удобно ввести понятие вариационного ряда.

Вариационный ряд – это упорядоченные данные, расположенные в порядке возрастания значения переменной, либо в порядке убывания.

Пример (вариационные ряды из отметок студента)

Исходный набор из отметок студента:

5, 4, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5

После упорядочения (в порядке возрастания) получим вариационный ряд:

2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5

В порядке убывания получим другой вариационный ряд:

5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 3, 3, 3, 2

Вариационный ряд – это упорядоченные данные, расположенные в порядке возрастания значения переменной, либо в порядке убывания. Ряд называется вариационным потому, что содержит варианты значений признака. Рассмотрим пример построения вариационных рядов из отметок студентов (исходные данные видны на рисунке выше). По нему построим два вариационных ряда. Первый ряд упорядочен по возрастанию, второй – по убыванию.

Теперь можем определить понятие медианы. Медиана (обозначается Me) это значение, соответствующее среднему (или серединному) элементу вариационного ряда. Понятие «серединный элемент» отличается для четного и нечетного количества значений переменной.

Заметим, что в случае, когда число элементов в наборе данных нечетно, медиана единственна. В случае же, когда число элементов четно, в качестве медианы можно взять любое число из отрезка с "серединными концами". Чаще же, в качестве медианы выбирают полусумму этих серединных элементов.

Для набора данных из n значений, если n нечетно, серединный элемент имеет номер $\frac{n+1}{2}$, а для четного значения n медиана находится как среднее арифметическое двух соседних серединных элементов с номерами $\frac{n}{2}$ и $\frac{n}{2}+1$. Медиана может быть определена для числовых и порядковых данных. Для каждого набора данных имеется только одна медиана.

Пример (вычисление медианы для отметок студента)

2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5

14 элементов

Медиана вычисляется как среднее значение 7 и 8 элементов (4 + 5)/2

Me = 4.5

Рассмотрим пример вычисления медианы для отметок студента. Возьмем упорядоченный по возрастанию вариационный ряд с отметками. В нем 14 элементов. Значит, медиана вычисляется как среднее значение 7-го и 8-го элементов и равна 4.5.

Еще один пример – вычисление медианы для показателей силы морского ветра по шкале Бофорта. Есть следующие наблюдения о силе морского ветра по шкале Бофорта, которые вы видите на экране. Сформируем на их основе вариационный ряд по возрастанию значения переменной. Количество элементов – 13. Значит 7-й по порядку элемент и есть медиана.

Подведем итоги. Мы рассмотрели 3 характеристики центральной тенденции. В следующей таблице указано, какие характеристики могут быть применимы к тем или иным шкалам.

Какая из этих характеристик лучше? Какую из них применять, если есть возможность выбора? На первый взгляд может показаться, что среднее – наи-

Пример (вычисление медианы для показателей силы морского ветра по шкале Бофорта)

0, 2, 2, 1, 1, 3, 3, 1, 1, 0, 0, 1, 2

Вариационный ряд по возрастанию

0, 0, 0, 1, 1, 1, (1), 1, 2, 2, 2, 3, 3

Me = 1 – тихий ветер

0 баллов штиль	4 балла умеренный ветер	8 баллов очень крепкий ветер
1 балл тихий ветер	5 баллов свежий ветер	9 баллов шторм
2 балла лёгкий ветер	6 баллов сильный ветер	10 баллов сильный шторм
3 балла слабый ветер	7 баллов крепкий ветер	11 баллов жестокий шторм
		12 баллов ураган

более емкая, широко известная и применяемая на практике характеристика. В плане известности, несомненно, да, но в плане полезности применения – далеко не всегда. Приведем хорошо известный пример на эту тему.

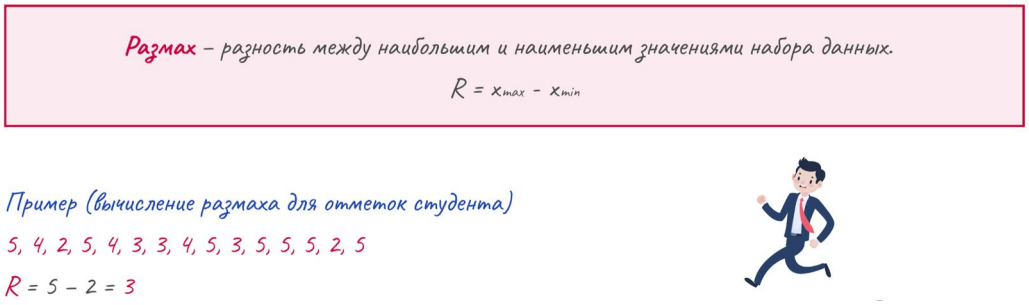
Характеристики центральной тенденции	Номинальные данные	Порядковые данные	Интервальные данные	Относительные данные
Мода	✓	✓	✓	✓
Медиана		✓	✓	✓
Среднее			✓	✓

В некоторой деревне проживает 50 жителей. Среди них 49 человек – сельские жители с месячным доходом 1 тыс. рублей, а один житель – зажиточный фермер с доходом 451 тыс. рублей. Вычислим средний доход жителей деревни. Он равен 10 тыс. рублей. Совершенно очевидно, что это число не отражает адекватно доход жителей деревни. В этом случае гораздо более рационально было бы использовать в качестве меры центральной тенденции моду или медиану (обе равны 1 тыс. рублей). Кроме того, в этом случае одного числа явно не хватает для описания доходов жителей в этой деревне.

1.1 Размах и квартильный размах

Мера центральной тенденции — всего лишь одно число, которое используется для описания типового значения из изучаемой выборки. Оно не дает представления о том, насколько разнообразны данные в выборке. Именно поэтому было придумано понятие размаха и квартильного (или межквартильного) размаха.

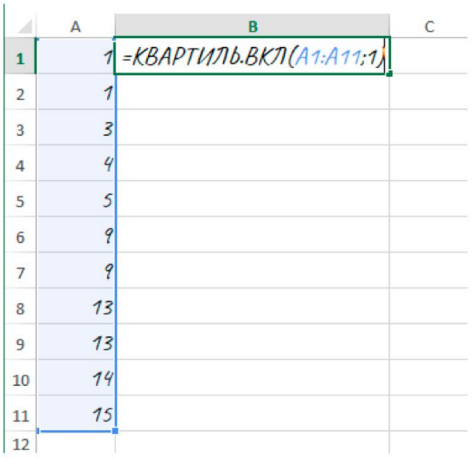
Размах – разность между наибольшим и наименьшим значениями набора данных. Для набора данных, представляющих отметки студента, который вы видите на экране, размах равен 3.



Следующая характеристика разброса данных в выборке – **квартильный (или межквартильный) размах**. Он основывается на понятии квартилей. Под квартилями понимаются значения Q_1, Q_2, Q_3 которые делят вариационный ряд на четыре равные части.

Второй квартиль Q_2 совпадает с медианой. Q_1 – это медиана для значений, которые меньше Q_2. Q_3 – это медиана для значений, которые больше Q_2. Существует несколько вариантов точного определения значения квартилей, которые могут немного отличаться. Например, при определении первого Q_1 и третьего квартилей Q_3 можно включать или, наоборот, исключать медиану (то есть слова меньше/больше понимать как строго больше/строго меньше

Квартили (Quartile)

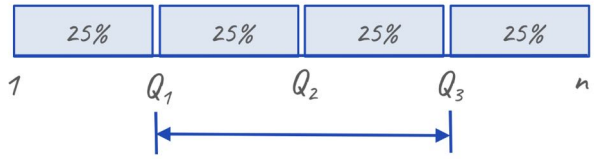


КВАРТИЛЬ.ИСКЛ() или QUARTILE.EXC()
КВАРТИЛЬ.ВКЛ() или QUARTILE.INC()

или не строго больше/не строго меньше). Именно поэтому в большинстве современных инструментов есть две версии функции квартиль: с исключением и включением медианы при определении первого и третьего квартилей. Называются такие функции: QUARTILE.EXC()/КВАРТИЛЬ.ИСКЛ() и QUARTILE.INC()/КВАРТИЛЬ.ВКЛ().

Размах квартилей – это разница между третьим и первым квартилем и вычисляется по формуле:

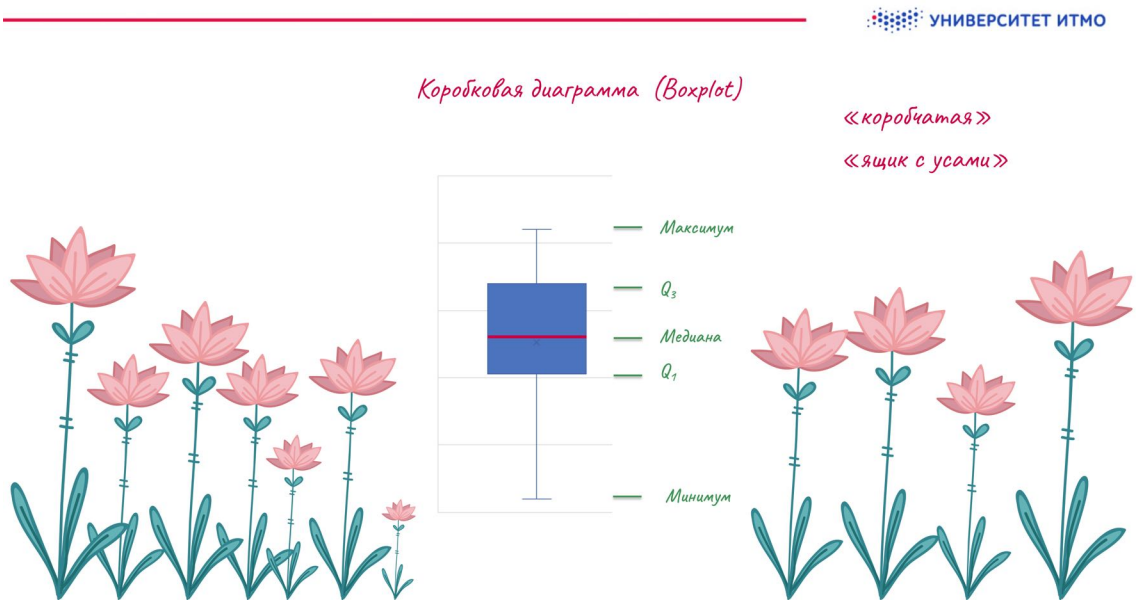
$$IQR = Q_3 - Q_1$$



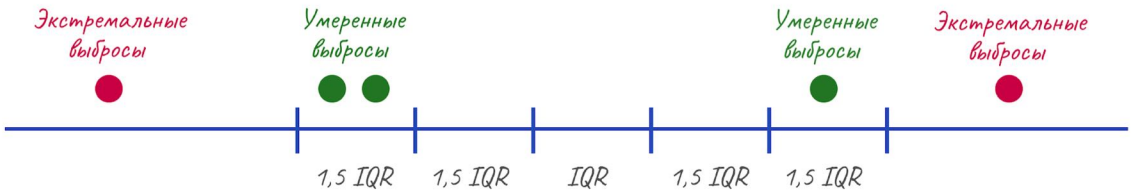
В чем принципиальная разница между размахом и квартильным размахом? Размах – очень простая и «грубая» мера вариации, т.к. при вычислении размаха используются только наименьшее и наибольшее значения переменной. При вычислении квартильного размаха игнорируются только крайние значения, расположенные за пределами первого и третьего квартилей. Между третьим и первым квартилем оказываются 50% всех данных.

При проведении разведочного анализа очень полезной оказывается, так называемая, **коробковая диаграмма**. Она имеет вид, изображенный на рисунке ниже, и может быть нарисована как в горизонтальном, так и в вертикальном виде. На ней отображены минимум, максимум и три квартиля. Это позволяет очень емко и выразительно отобразить основные значения данных.

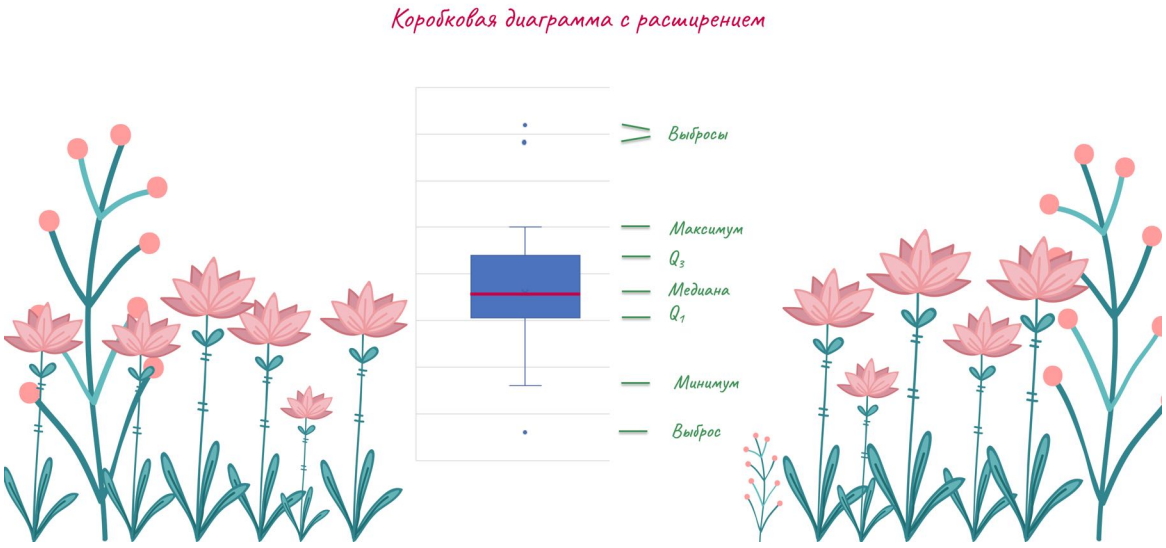
Кроме того, квартили могут использоваться для определения, так называемых, выбросов, т.е. значений, которые слишком отличаются от остальных.



Классифицируют два вида выбросов: **умеренные** и **экстремальные**. **Умеренными** называют выбросы, которые удалены ниже первой квартили или выше третьей от $1.5IQR$, но не более, чем на $3IQR$. **Экстремальные выбросы** удалены ниже первой квартили или выше третьей более, чем на $3IQR$. Схема определения выбросов приведена на рисунке. Определение выбросов крайне важно на этапе подготовки данных и позволяет избавиться значений, достоверность которых сомнительна.



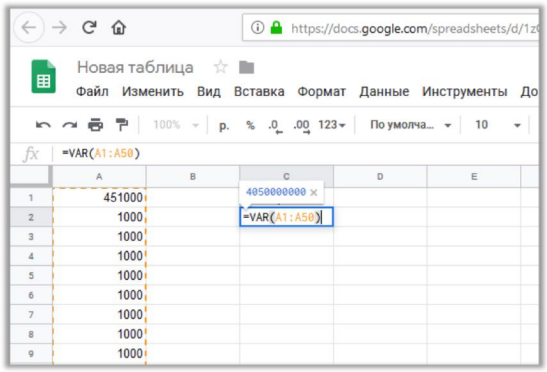
Кроме того, на основании этих данных рисуется, так называемая, коробковая диаграмма с расширением, на которой отображены выбросы. Рассчитывается она в два этапа: на первом определяются квартили и по ним – выбросы (они будут отображены на диаграмме в виде точек), а затем из данных исключаются выбросы и заново пересчитываются минимум, максимум и квартили и отображаются в виде обычной коробковой диаграммы. Пример такой коробковой диаграммы с расширением вы можете видеть на рисунке.

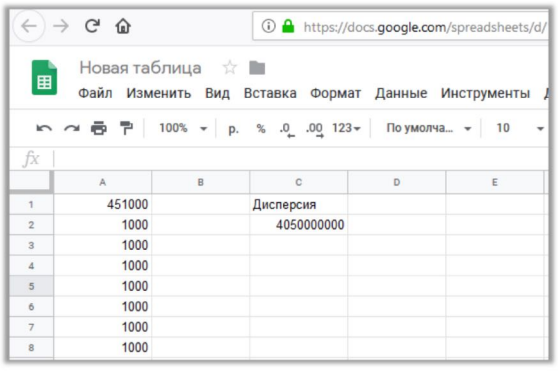


Еще одно очень полезное статистическое понятие – **дисперсия**. Дисперсия для набора данных или выборки – это среднее арифметическое квадратов отклонений значений от их среднего. Значение дисперсии может быть вычислено явно по упомянутой выше формуле или с помощью любого подходящего инструментария, в котором такая функция присутствует.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

В GOOGLE таблицах такая функция присутствует и называется ДИСП() или VAR(). Укажем для этой функции исходную выборку – доходы жителей упомянутой ранее деревни. И получим следующий результат, который вы можете видеть на экране.

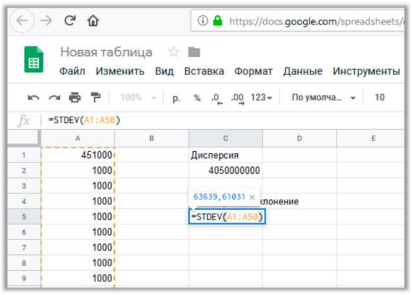




С понятием дисперсии тесно связана еще одна описательная статистика – стандартное отклонение. **Стандартное отклонение** – это квадратный корень из дисперсии выборки. Вычисляется по формуле:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Присутствует практически во всех инструментах. В GOOGLE таблицах ему соответствует функция СТАНДОТКЛОН()/STDEV(). На рисунке вы можете видеть вычисление стандартного отклонения для доходов жителей деревни.

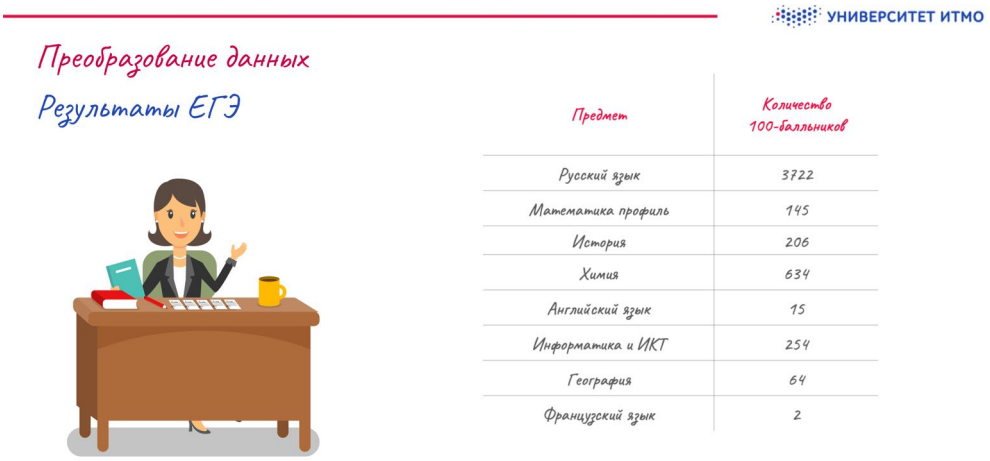


СТАНДОТКЛОН() или STDEV()

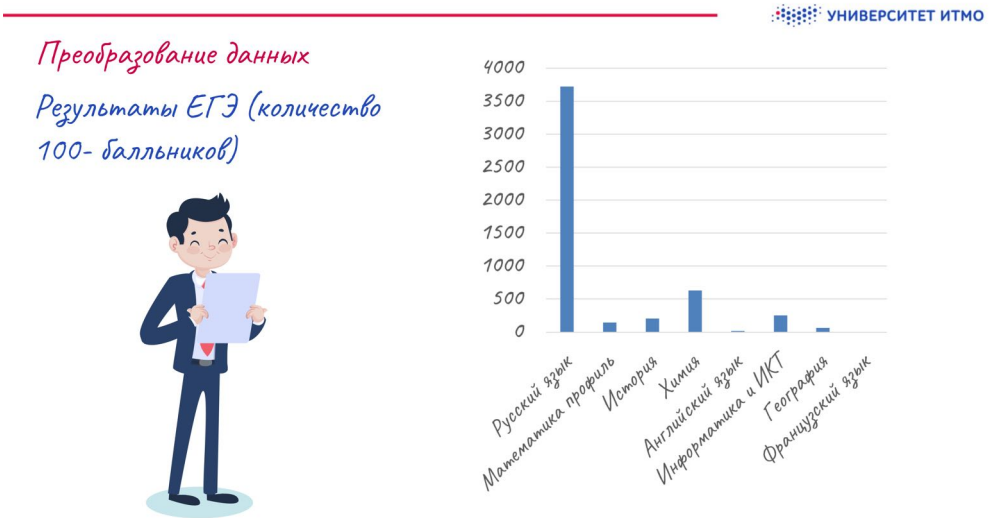
Итак, в данном разделе мы рассмотрели основные описательные статистики, которые могут оказаться полезными на этапе разведочного анализа данных.

2 Преобразование данных

Преобразование данных – одна из распространенных процедур предварительной обработки данных, способная продемонстрировать характерные особенности, скрытые в данных и не видимые в их первоначальной форме.



Попытаемся аргументировать необходимость проведения преобразований на конкретном примере. В нашем распоряжении есть агрегированные данные о результатах ЕГЭ за 2018 год. Нет никакого сомнения, что они достоверны. Это данные из официальных релизов Рособрнадзора.

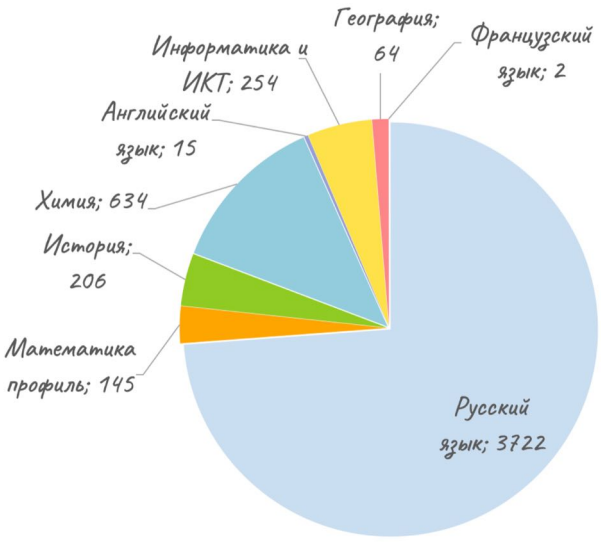


В таблице слишком много чисел, которые трудно охватить и сделать общие выводы. Попытаемся отобразить в виде простейших графиков количество 100-балльников по различным предметам. Данные по своему содержанию соответствуют агрегированным значениям из различных категорий.

Значит им подойдет визуализация в виде столбчатой или круговой диаграммы. Вот как выглядят наши данные на столбчатом графике.

Еще раз подчеркнем, что данные достоверны. Это агрегированные данные, которые ровно такие, как они есть. Но совершенно очевидно, что визуализировать их таким образом нельзя, так как некоторые значения (например, соответствующие категориям Английский язык, География и Французский язык) отображаются неадекватно – они почти не видны на графике.

Попробуем отобразить эти же данные в виде круговой диаграммы. Воз-



можно, стало немного лучше. География стала выглядеть более убедительно. Однако Английский и Французский языки, по-прежнему, практически не видны на диаграмме. Причина, по которой указанные значения не видны – большой разброс значений среди агрегированных данных. Что делать с такими данными и как их визуализировать? Возможное решение – преобразование данных таким образом, чтобы разброс значений уменьшился, а сами данные стали, как минимум соизмеримыми. Существует много различных методов преобразования. Рассмотрим наиболее традиционные и обсудим, как эта трансформация влияет на визуализацию.

2.1 Распространенные преобразования

В таблице на ниже приведены наиболее распространенные способы преобразования и особенности их использования.

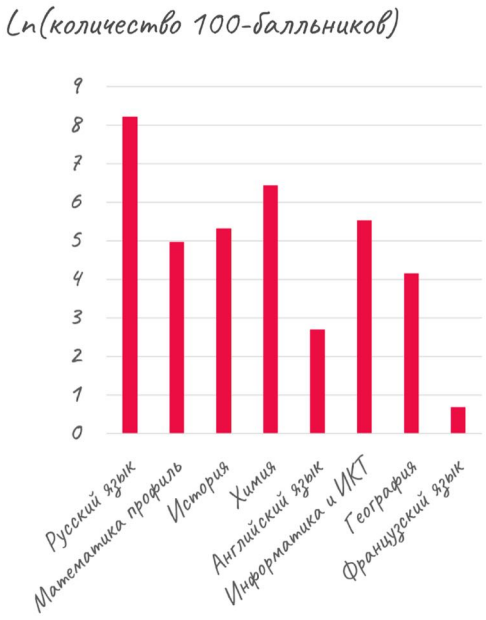
Так, например, логарифм натуральный или десятичный хорошо подходит для преобразования данных, сохраняет порядок среди значений, но не уместен при наличии нулевых значений в исходных данных. Преобразование квадратного корня также сохраняет порядок между значениями, уместно при нулевых значениях, но не уместно при наличии отрицательных. Преобразование по формуле обратной дроби меняет порядок значений (что иногда может оказаться уместным), но не умеет работать с нулевыми значениями.

Преобразование	Не подходит для
$\ln(x)$	нулевых значений
$\log_{10}(x)$	нулевых значений
$\sqrt{x}$	отрицательных значений
x^2	отрицательных значений
$1/x$	нулевых значений

Логарифмическое преобразование. Рассмотрим, как применяются преобразования на примере логарифмического преобразования. Чтобы осуществить логарифмическое преобразование, необходимо вычислить логарифм каждого значения в наборе данных и использовать эти преобразованные данные вместо исходных. Логарифмические преобразования оказывают

Предмет	Количество 100-балльников	Ln(количество 100-балльников)
Русский язык	3722	8.22
Математика профиль	145	4.98
История	206	5.33
Химия	634	6.45
Английский язык	15	2.71
Информатика и ИКТ	254	5.54
География	64	4.16
Французский язык	2	0.69

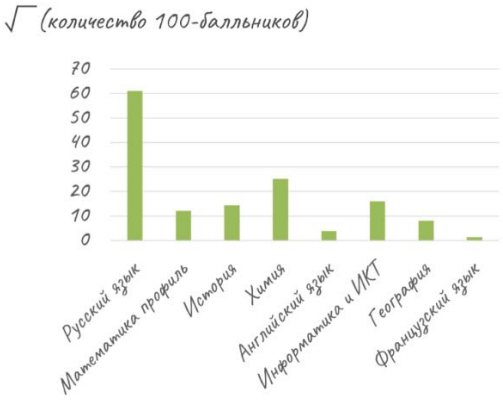
существенный эффект на форму распределения. На рисунке представлена столбчатая диаграмма о 100-балльниках ЕГЭ после применения натурального логарифмического преобразования.



Преобразование квадратного корня. Преобразование квадратного корня оказывает более умеренный эффект на форму распределения.

Следующий график показывает столбчатую диаграмму о 100-балльниках ЕГЭ после применения преобразования квадратного корня.

Предмет	Количество 100-балльников	√(количества 100-балльников)
Русский язык	3722	61.01
Математика профиль	145	12.04
История	206	14.35
Химия	634	25.18
Английский язык	15	3.87
Информатика и ИКТ	254	15.94
География	64	8.00
Французский язык	2	1.41



2.2 Выбор подходящего преобразования

Возможных преобразований много. Как выбрать подходящее преобразование? Ответ на этот вопрос не очевиден, хотя формальные статистические методы для выбора преобразования существуют. Если не вникать в эти теории, то общая стратегия выбора преобразования заключается в том, чтобы четко определить цель преобразования (например, визуализация определенного типа, сохранение или, наоборот, разворот упорядоченности), а затем применить наиболее используемые преобразования, такие как логарифмы, квадратный корень, квадрат, обратная дробь и выбрать лучший метод, исхо-

дя из цели и полученных результатов.

2.3 Единицы измерения и обратные преобразования

Поскольку методы преобразования включают в себя применение к исходным данным математических функций, то необходимо обратить внимание на изменение единиц измерения данных. Например, при применении логарифмической функции к переменной численности 100-балльников, единицей измерения становится логарифм численности.

Значит, при представлении данных на графиках и диаграммах надо явно указывать, какие именно преобразования были проведены, и в каких единицах измерения отображены данные. Если преобразованные данные использовались для вычисления статистик, то надо не забыть провести обратное преобразование, чтобы представить результат в начальных единицах измерения. Например, если было применено преобразование квадратного корня, необходимо совершить обратное преобразование, и возвести конечный результат в квадрат.

3 Нормировка данных

Нормировка данных – еще одна процедура возможной предварительной обработки данных. Назначение нормировки – обеспечение возможности для сравнения, агрегации и, возможно, визуализации значений нескольких переменных из различных шкал. Для некоторых алгоритмов машинного обучения (и не только) нормировка переменных является необходимым условием.

Попытаемся аргументировать целесообразность нормировки на очень простом конкретном примере. В распоряжении педагогического совета школы есть одна путевка в Артек и сведения об учащихся, представленные на слайде. Будем условно считать, что все грамоты приблизительно одного по-

	Средняя отметка за период обучения	Количество грамот за участие в художественных конкурсах	Количество грамот за участие в интеллектуальных конкурсах	Количество грамот за участие в спортивных мероприятиях
Иван	4.5	5	2	8
Александра	4.9	0	4	4
Семен	4	2	3	10
Екатерина	4.5	5	3	0
Иннокентий	4.2	2	6	9
Анна	5	2	5	0

рядка и среди них нет каких-то невероятно выдающихся. Перед педсоветом стоит задача найти (а в дальнейшем опубликовать) формальный критерий, по которому будет отобран претендент на единственную путевку в Артек. Как найти этот критерий? Задача состоит в том, чтобы каждому школьнику сопоставить одно число, которое будет представлять все его достижения, а затем на основании этих чисел составить рейтинг школьников и выявить первого претендента. Если бы все переменные и отметки измерялись в одинаковых шкалах и единицах измерения, можно было бы предложить сложить все значения, но это очень грубый подход, так как спортивные грамоты выдаются гораздо чаще других, и спортивные достижения тут же перекроют все остальные. Выход – нормировка значений переменных, а затем вычисление на их основе итогового критерия.

Почему нужна нормировка показателей? Обычно выраженность некоторого качества описывают числом. Переменная x меняется от некоторого минимального значения x_{min} (отражающего отсутствие качества) до некоторого максимального значения x_{max} (высшая степень проявления качества). Наличие критерия качества позволяет решать проблему сравнения двух объектов, но только по этому показателю. Но при этом надо помнить, в каких пределах меняется показатель. А диапазоны разброса значений и единицы измерений для разных переменных — самые разнообразные... Кроме того иногда необходимо оценивать, насколько близко конкретное значение к краям диапазона или к его середине. Если же речь идет о сравнении или агрегировании по различным показателям — дело обстоит совсем плохо. А ведь именно показатель качества интерпретируется как **степень выраженности** качества. А степени выраженности сравнивать и агрегировать можно и нужно! Но для этого показатели следует привести к одной шкале так, чтобы минимальное и максимальное значения для различных переменных совпадали. Такое преобразование и называется **нормировкой**. После этого преобразования можно сравнивать и агрегировать разнообразные показатели, полученные различными методиками.

3.1 Классы числовых показателей

При всем разнообразии числовых характеристик объектов из них можно выделить два широких класса:

- **униполярные**, выражающие только степень наличия некоторого качества или количества (например, интенсивный цвет, очень хорошая отметка или количество чего-либо);
- **биполярные**, отражающие не только степень наличия качества, но и его «направленность».

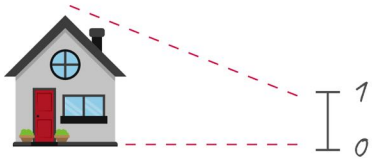
Методы нормировки различаются для этих классов. Рассмотрим последовательно некоторые из них.

3.2 Нормировка униполярных показателей

Обычно униполярные показатели нормируются в диапазоне от 0 до 1. В качестве функции нормировки может выступать любая непрерывная возрастающая функция $y = f(x)$ с минимальным значением – 0 и максимальным значением – 1.

Функция нормировки для униполярного показателя

$$y(x_{min}) = 0; \quad y(x_{max}) = 1; \quad \frac{dy}{dx} > 0$$



Рассмотрим возможные варианты такой функции, безусловно, обладающих упомянутыми выше свойствами. Существует два возможных варианта нормировки, **экспоненциальная**

$$y(x) = 1 - \exp\left(1 - \frac{x}{x_{min}}\right)$$

или **линейная**

$$y(x) = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Нормировка униполярных показателей

Вариант I

Экспоненциальная нормировка

$$y(x) = 1 - exp(1 - \frac{x}{x_{min}})$$

Вариант II

Линейная нормировка

$$y(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Надо заметить, что линейная функция в силу своей простоты используется чаще других. Достоинством экспоненциальной функции считается то,

что она равномернее распределяет исходные значения по диапазону от 0 до 1. И более того, небольшие модификации этой формулы с легкостью позволяют усилить эту равномерность распределения в конкретных случаях.

3.3 Нормировка биполярных показателей

Биполярные показатели обычно нормируются в диапазоне от -1 до 1. В качестве функции нормировки может выступать любая непрерывная возрастающая функция $y = f(x)$ с минимальным значением минус 1 и максимальным значением плюс 1. Такой линейной функции, основанной на минимальных и максимальных значениях, может быть:

$$y(x) = \frac{2x - (x_{max} + x_{min})}{x_{max} - x_{min}}.$$

Разумеется, есть и другие возможные варианты нормировки и некоторые из них не являются линейными и порою привязаны к специфике предметной области, в которой проводится нормировка показателей, тем не менее, в большинстве случаев такое преобразование является вполне достаточным для последующего анализа. Теперь, вооружившись полученными знаниями,

вернемся к нашему примеру со школьниками.

3.4 Какие показатели у учащихся?

Рассмотрим все показатели нашего примера с учащимися. Все они – униполярные. Действительно, средняя отметка – униполярный показатель, который отражает однонаправленное качество успехов в обучении, как правило, измеряется от 1 до 5. Количество грамот за участие в художественных конкурсах – униполярный показатель (представлен в виде положительного целого). Количество грамот за спортивные достижения – униполярный показатель (представлен в виде положительного целого). Количество грамот за участие в интеллектуальных конкурсах – униполярный показатель (представлен в виде положительного целого). Это означает, что при нормировке данных мы можем воспользоваться любой из нормировок для униполярных

Униполярный показатель (отражает качество успехов в обучении, измеряется от 1 до 5)

Униполярные показатели (отражают количественные успехи, представлены в виде положительных целых чисел)

	Средняя отметка за период обучения	Количество грамот за участие в художественных конкурсах	Количество грамот за участие в интеллектуальных конкурсах	Количество грамот за участие в спортивных мероприятиях
Иван	4.5	5	2	8
Александра	4.9	0	4	4
Семен	4	2	3	10
Екатерина	4.5	5	3	0
Иннокентий	4.2	2	6	9
Анна	5	2	5	0

Пример нормировки показателя

$x_{min} = 0;$

$x_{max} = 5;$

	Количество грамот за участие в художественных конкурсах	
	до нормировки	после нормировки
Иван	5	1.00
Александра	0	0.00
Семен	2	0.40
Екатерина	5	1.00
Иннокентий	2	0.40
Анна	2	0.40

показателей. Для простоты возьмем линейную нормировку. Рассмотрим, к примеру, как будет выполнена нормировка показателя «Количество грамот за участие в художественных конкурсах». Для начала определим минимальные и максимальные значения для этого показателя. Они равны соответственно – 0 и 5. Подставим эти значения в формулу линейной нормировки значения для показателя. Результаты вы можете видеть на экране.

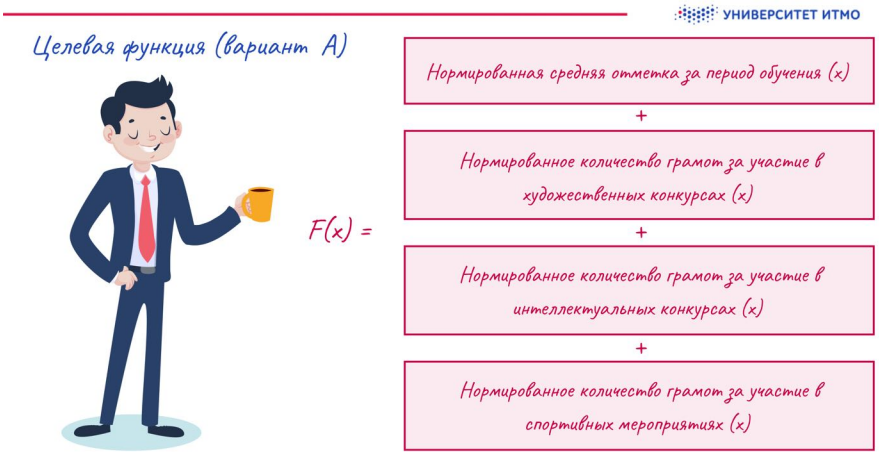
Воспользуемся формулой униполярного линейного преобразования для всех показателей и нормируем все исходные данные (причем нормировать надо каждый показатель по отдельности). Итак, у нас есть нормированные значения по каждому показателю. Что дальше?

Теперь необходимо задать, так называемую, целевую функцию на основе нормированных значений (соответствующих ученикам). Что из себя представляет целевая функция? Это математическое выражение некоторого критерия качества объекта (процесса, решения). Целевая функция задается для того, чтобы вместо большого количества качественных параметров для каждого изучаемого объекта получить один, а затем на основе максимального

Показатели учащихся после нормировки

	Средняя отметка за период обучения	Количество грамот за участие в художественных конкурсах	Количество грамот за участие в интеллектуальных конкурсах	Количество грамот за участие в спортивных мероприятиях
Иван	0.50	1.00	0.00	0.80
Александра	0.90	0.00	0.50	0.40
Семен	0.00	0.40	0.25	1.00
Екатерина	0.50	1.00	0.25	0.00
Иннокентий	0.20	0.40	1.00	0.90
Анна	1.00	0.40	0.75	0.00

(или минимального) значения функции определить объект, на котором достигается соответствующий экстремум. Так какое же значение функции нужно использовать? Максимум? Минимум? Что именно – зависит от специфики поставленной задачи и вида самой функции.



Например, если эта функция отражает суммарные положительные качества ученика, то это, наверняка, максимум. А если это суммарная стоимость затрат для выполнения какой-то задачи, то логичнее использовать минимум. В нашем случае в качестве целевой функции можно использовать сумму нормированных значений, так как каждое из значений отражает какие-то положительные качественные характеристики ученика, а лучшим значением считать максимум такой функции. Ученик, для которого функция выдаст максимальное значение, будет считаться лучшим. В нашем распоряжении есть, как минимум, два возможных варианта, чтобы увидеть результат вычисления такой функции: мы можем добавить к таблице еще один столбец, в котором будет вычислена сумма нормированных показателей и найти учащегося с максимальным значением целевой функции или использовать замечательное средство для визуализации – гистограмму с накоплением. Этот тип диаграмм есть в большинстве популярных инструментов визуализации. Особенностью этой диаграммы является то, что она сама суммирует показатели и наша

задача их всего лишь правильно задать и найти столбец с максимальным накопленным значением. Продемонстрируем оба варианта.

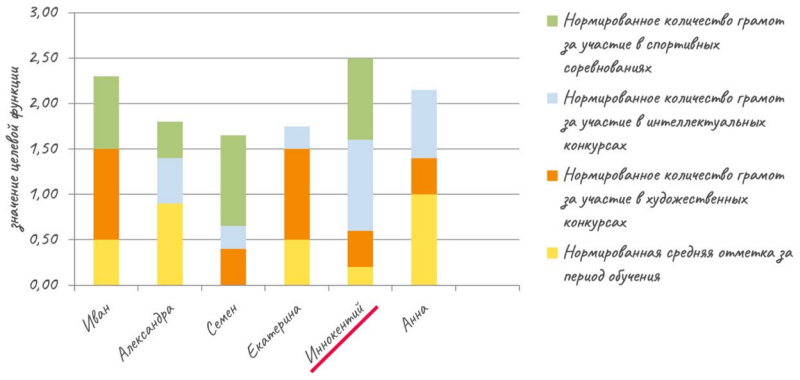
Таблица с явно вычисленной целевой функцией

	Нормированная средняя отметка за период обучения	Нормированное количество грамот за участие в художественных конкурсах	Нормированное количество грамот за участие в интеллектуальных конкурсах	Нормированное количество грамот за участие в спортивных мероприятиях	Значение целевой функции (вариант А)
Иван	0.50	1.00	0.00	0.80	2.30
Александра	0.90	0.00	0.50	0.40	1.80
Семен	0.00	0.40	0.25	1.00	1.65
Екатерина	0.50	1.00	0.25	0.00	1.75
Иннокентий	0.20	0.40	1.00	0.90	2.50
Анна	1.00	0.40	0.75	0.00	2.15

Итак, на рисунке ниже представлена таблица, в которой в качестве последнего столбца добавлено значение целевой функции (вариант А – сумма всех нормированных показателей). Легко видеть, что максимальное значение целевой функции соответствует Иннокентию. Значит Иннокентий – победитель! Рассмотрим второй возможный вариант определения победителя – гистограмму с накоплением.

Построим гистограмму с накоплением на основе нормированных значений и увидим абсолютно аналогичный результат. Именно Иннокентию соответствует столбец с накопленным значением максимальной высоты. Все, о чем нужно позаботиться в данном случае, это – правильно задать тип диаграммы и значения исходных данных.

Гистограмма с накоплением
Значение целевой функции
(вариант А)



На этом можно было бы остановиться, но оказалось, что педагогический совет настаивает, чтобы вдвое усилить значимость нормированного балла за успеваемость в школе, т.е. если все остальные нормированные значения войдут в целевую функцию с коэффициентом 1, то у нормированного среднего

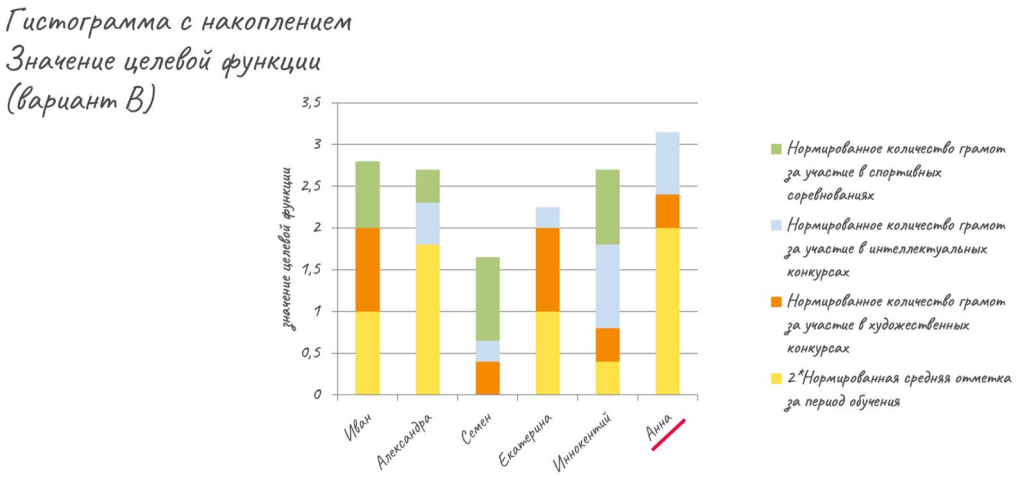
балла будет коэффициент значимости 2 (такие коэффициенты принято называть весовыми коэффициентами). Итоговая формула для целевой функции в этом случае отображена на рисунке ниже.

Попробуем определить победителя с новой целевой функцией. На рисунке представлена таблица, в которой в качестве целевой функции используется вариант B. Легко видеть, что максимальное значение целевой функции (2.833) соответствует Анне.

Таблица с явно вычисленной целевой функцией

	Нормированная средняя отметка за период обучения • 2	Нормированное количество грамот за участие в художественных конкурсах	Нормированное количество грамот за участие в интеллектуальных конкурсах	Нормированное количество грамот за участие в спортивных мероприятиях	Значение целевой функции (вариант B)
Иван	1.00	1.00	0.00	0.80	2.8
Александра	1.80	0.00	0.50	0.40	2.7
Семен	0.00	0.40	0.25	1.00	1.65
Екатерина	1.00	1.00	0.25	0.00	2.25
Иннокентий	0.40	0.40	1.00	0.90	2.7
Анна	2.00	0.40	0.75	0.00	3.15

Значит, в этом случае победителем является Анна! Аналогичный вариант мы можем увидеть и на гистограмме с накоплением.



Мы не можем изменить манеру накопления в гистограмме, но мы можем удвоить значения показателя за успеваемость в исходных данных для диаграммы и на их основе построить обычную диаграмму с накоплением. Как и ожидалось, победителем оказалась Анна.

Построение целевой функции на основе нормированных показателей, подбор подходящих весовых коэффициентов в общем случае весьма непростая задача, которая выходит за рамки нашей лекции. Однако к этой теме мы непременно вернемся в рамках курса по Машинному обучению.

4 Целевая функция

Иногда, работая с объектами реального мира, нам нужно сравнивать объекты между собой, и находить «лучший» вариант, решая ту или иную прикладную задачу. Если объект обладает только одним признаком, то задача кажется элементарной: найти самую дешевую квартиру, автомобиль с самым большим объемом багажника или самую высокую точку планеты над уровнем моря. Но если у объектов много признаков, которые надо взять в расчет, то как найти тот самый «лучший» из них? Если мы работаем с числовыми данными, то их можно нормировать, чтобы выровнять порядок возможных значений и привести диапазоны значений признаков к некоторым фиксированным границам. На основании нормированных данных принято строить, так называемые, целевые функции, которые используются для решения тех или иных прикладных задач. Как в общем случае можно дать определение целевой функции?

Целевая функция – числовая функция от нескольких переменных (параметров), подлежащая оптимизации в целях решения некоторой оптимизационной задачи.

Что означает слово **оптимизация** в таком контексте? Это поиск тех возможных значений параметров, при которых функция принимает максимум или минимум (зависит от решаемой прикладной задачи).

Решение прикладных задач часто сводится к решению задач оптимизации. Причем при решении последних, обычно придерживаются следующего плана:

1. Нормируются значения входных переменных (признаков).
2. Исходя из целей задачи, строится целевая функция (функция, подлежащая оптимизации).
3. Ищется наибольшее (или наименьшее) значение целевой функции на рассматриваемом множестве.

В последнем пункте предложенного алгоритма часто вместо самого значения целевой функции, оказывается интересным значение аргумента, при котором это значение достигается. Такая точка часто называется оптимальной точкой целевой функции, или точкой глобального максимума (минимума).

$$F(X) = a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n,$$

где x_i — нормированные параметры, a_i – коэффициенты (веса).

В простейших случаях целевая функция – это сумма числовых нормированных параметров (иногда с некоторыми числовыми коэффициентами, которые принято называть весами). Нормировка позволяет «подравнять» значения разных признаков, а коэффициенты (или веса) целевой функции позволяют, наоборот, сделать тот или иной признак более или менее значимым. Но далеко не всегда целевая функция задается таким простым образом. Кроме того, следует заметить, что само построение целевой функции в большинстве случаев отражает субъективное мнение исследователя, решающего конкретную прикладную задачу. Фактор субъективности проявляется особенно явно при построении целевой функции, если признаки не являются числовыми (например, номинальными, порядковыми и т.п.). Именно исследователь должен решить, как такие данные нужно преобразовывать к численному виду, а уж затем строить подходящую целевую функцию. Отметим, что в настоящее время для преобразования значений существует множество приемов и готовых пакетов, встроенных в различные программные среды.

Одним из наиболее универсальных приемов при преобразовании номинальных данных к числовому виду является следующий: каждому номинальному значению сопоставляется дополнительный признак с двумя возможными значениями. Например, если переменная «цвет» может принимать 3 возможных значения: «голубой», «красный» и «белый», то можно вместо признака «цвет» ввести три дополнительных признака: «голубой?», «красный?»,

Номинальные значения

id объекта	Цвет
1	голубой
2	красный
3	белый

Числовые значения

id объекта	Голубой?	Красный?	Белый?
1	1	0	0
2	0	1	0
3	0	0	1

«белый?», каждый из которых при описании конкретного объекта может принимать значения 0 или 1.

С порядковыми данными принято поступать несколько иначе. Как известно, порядковые данные можно упорядочить (поэтому они так и называются). А, следовательно, им можно сопоставить числовые значения, которые отразят их порядок. Например, если рассмотреть ответы на некоторый вопрос: «За», «Против», «Категорически против», «Безоговорочно поддерживаю», «Безразлично», то их можно упорядочить относительно числовой шкалы, отражающей «меру согласия», и сопоставить числовые значения, которые сохранят этот порядок. Например, в биполярном и униполярном варианте значения могли бы выглядеть так:

Биполярный вариант

Категорически против	2
Против	1
Безразлично	0
За	-1
Безоговорочно поддерживаю	-2

Униполярный вариант

Категорически против	0
Против	1
Безразлично	2
За	3
Безоговорочно поддерживаю	4

Попробуем рассмотреть конкретный пример построения целевой функции для выбора лучшего автомобиля в аренду. Исходные данные – сведения об автомобилях каршеринговой компании. Будем полагать, что выбирать марку автомобиля не требуется, так как все автомобили одной марки, но вот опции у этих автомобилей могут различаться. Известно, что:

- некоторые автомобили работают на дизельном топливе, некоторые — на бензине;
- автомобили имеют одну из трех возможных коробок передач: автоматическую, механическую или роботизированную;

- в некоторых автомобилях есть опция подогрев руля, а в некоторых — нет;
- автомобили имеют различный объем багажника;
- различаются также годы выпуска автомобилей и стоимость аренды за один день.

Как выбрать лучший автомобиль? Начнем с того, что представим все данные в числовом виде, затем нормируем эти данные, потом построим подходящую целевую функцию и на основе ее экстремального значения (максимум или минимум) выберем лучший автомобиль.

Приведем список доступных автомобилей в табличном виде:

Номер варианта	Топливо	Коробка передач	Подогрев руля	Объем багажника (л)	Год выпуска	Стоимость аренды (руб)
1	дизель	автомат	да	270	2015	2000
2	бензин	автомат	да	700	2014	2100
3	дизель	механика	нет	370	2018	2150
4	бензин	автомат	да	750	2018	2200
5	дизель	автомат	да	370	2016	1900
6	дизель	автомат	нет	700	2017	2500
7	бензин	механика	нет	350	2019	2300
8	дизель	автомат	да	700	2017	2400
9	бензин	автомат	да	750	2017	2600
10	бензин	автомат	нет	800	2017	2700
11	бензин	роботизирована	нет	850	2019	2800

Итак, приступим к нормировке данных. Рассмотрим опцию – подогрев руля. С точки зрения данных — это дихотомические данные, у которых есть два возможных значения — да и нет. Эти данные содержательно можно представить в виде чисел 0 и 1, с которыми дальше можно будет строить целевую функцию. «Подогрев руля» — отличная опция. Поэтому естественно представить значение «да» как 1, а «нет» как 0. Замечательно то, что данные сразу же оказались нормированными и не потребуют дальнейшего преобразования.

Подогрев руля (исходные значения)	Подогрев руля (числовые значения)
да	1
нет	0

Как преобразовать данные о коробке передач? С точки зрения водителя (именно здесь и проявляется субъективность подхода исследователя) эти данные можно представить как порядковые, а, следовательно, числовые. Причем с учетом предпочтений типичного водителя, которому больше нравится автоматическая коробка передач, числа можно сопоставить так:

Коробка передач (исходные значения)	Коробка передач (числовые значения)
автомат	3
роботизирована	2
механика	1

Обратите внимание, что максимальное значение опять сопоставлено «лучшему» значению. Но этого преобразования явно недостаточно, чтобы строить в дальнейшем целевую функцию. Нужно нормировать полученные значения. Используем классическую линейную нормировку и получим такие данные:

Коробка передач (исходные значения)	Коробка передач (числовые значения)	Коробка передач (нормированные числовые значения)
автомат	3	1
роботизирована	2	0,5
механика	1	0

Следующая опция – «вид топлива». С точки зрения данных – это номинальные данные, у которых есть два возможных значения: дизельное топливо и бензин. То есть, это опять дихотомические данные, которые можно представить в виде чисел 0 и 1. Многие водители предпочитают дизельное топливо (опять субъективный фактор) и поэтому, имеет смысл сопоставить дизельному топливу значение 1, а бензину – 0. Данные опять получились нормированными и не требуют дальнейшего преобразования:

Вид топлива (исходные значения)	Вид топлива (числовые значения)
дизель	1
бензин	0

Разберемся с опцией – «объем багажника». Данные числовые, и, следовательно, достаточно всего лишь провести нормировку этих данных. Результат линейной нормировки представлен в таблице.

Номер варианта	Объем багажника (до нормировки)	Объем багажника (после нормировки)
1	270	0,00
2	700	0,74
3	370	0,17
4	750	0,83
5	370	0,17
6	700	0,74
7	350	0,14
8	700	0,74
9	750	0,83
10	800	0,91
11	850	1,00

Следующая опция – «год выпуска автомобиля». Данные опять числовые и, следовательно, достаточно всего лишь провести нормировку этих данных. Результат линейной нормировки представлен в таблице.

Номер варианта	Год выпуска (до нормировки)	Год выпуска (после нормировки)
1	2015	0,20
2	2014	0,00
3	2018	0,80
4	2018	0,80
5	2016	0,40
6	2017	0,60
7	2019	1,00
8	2017	0,60
9	2017	0,60
10	2017	0,60
11	2019	1,00

И, наконец, последняя опция – «стоимость аренды». Данные опять числовые и, следовательно, достаточно всего лишь провести стандартную нормировку. Результат линейной нормировки представлен в таблице.

Итак, мы нормировали все данные об автомобилях, и остается построить целевую функцию F, найти ее максимум (или минимум) и определить,

Номер варианта	Стоимость аренды (до нормировки)	Стоимость аренды (после нормировки)
1	2000	0,11
2	2100	0,22
3	2150	0,28
4	2200	0,33
5	1900	0,00
6	2500	0,67
7	2300	0,44
8	2400	0,56
9	2600	0,78
10	2700	0,89
11	2800	1,00

для какого автомобиля целевая функция его достигает. Как уже говорилось выше, построение целевой функции – это субъективный процесс, отражающий мнение аналитика, который строит эту целевую функцию. Поэтому очень важно при построении целевых функций использовать мнения экспертов, которые разбираются в предметной области и могут аргументировать важность тех или иных признаков. В случае с выбором автомобиля попробуем сами выступить в роли экспертов и представим следующие требования, которые должны быть учтены при составлении целевой функции:

1. Все параметры, которые могут повлиять на выбор автомобиля, для нас одинаково важны (значит, никакие коэффициенты в целевой функции не понадобятся).

2. Если два автомобиля различаются только видом топлива, то предпочтем автомобиль с дизельным топливом (ему соответствует максимальное нормированное значение).

3. Если автомобили различаются только коробкой передач, то предпочтем автомобиль с коробкой передач, которой соответствует наибольшее нормированное значение.

4. Если автомобили различаются только наличием/отсутствием опции «подогрев руля», то предпочтем автомобиль с наличием опции (ему соответствует максимальное нормированное значение).

5. Если автомобили различаются объемом багажника, то предпочтем автомобиль с наибольшим объемом багажника (ему соответствует макси-

мальное нормированное значение).

6. Если автомобили различаются годом выпуска, то предпочтем автомобиль с максимальным годом выпуска (ему соответствует максимальное нормированное значение).

7. Если автомобили различаются стоимостью аренды, то предпочтем автомобиль с наименьшей стоимостью (ему соответствует минимальное нормированное значение).

Все эти экспертные соображения могут быть объединены в следующую целевую функцию, которая должна достигнуть максимума при выборе лучшего автомобиля:

$$F(X) = NF(X) + NB(X) + NH(X) + NV(X) + NY(X) + (1 - NP(X)),$$

где

- $NF(X)$ – нормированное значение, соответствующее виду топлива;
- $NB(X)$ – нормированное значение, соответствующее коробке передач;
- $NH(X)$ – нормированное значение, соответствующее опции подогрев руля;
- $NV(X)$ – нормированное значение, соответствующее объему багажника;
- $NY(X)$ – нормированное значение, соответствующее году выпуска автомобиля;
- $NP(X)$ – нормированное значение, соответствующее стоимости аренды автомобиля.

Обратите внимание на то, что при учете такого параметра, как стоимость аренды, в качестве слагаемого в целевой функции использовалось не само нормированное значение, а выражение $(1 - NP(X))$. Таким образом, мы пытаемся учесть тот факт, что при прочих равных условиях предпочтем автомобиль с минимальной стоимостью. В следующей таблице приведены значения целевой функции для всех автомобилей, представленных в исходных данных. Нам повезло и вариантов не так много. Нам удалось явным образом рассчитать значение целевой функции для каждого возможного варианта автомобиля. В таблице видно, что максимальное значение целевой функции соответствует варианту 8. Значит это и есть лучший автомобиль! Разумеется, в более сложных случаях не всегда удается явно просчитать все возможные

Номер варианта	Значения целевой функции F(X)
1	4,09
2	3,52
3	2,69
4	4,29
5	4,57
6	3,67
7	1,69
8	4,79
9	3,65
10	2,62
11	2,50

значения целевой функции и для поиска лучших значений существуют специальные математические методы. Но мы не будем говорить о них сейчас. О них вы сможете узнать в курсах по Машинному обучению.