Корреляция

4

Откуда Netflix известно, какие фильмы мне нравятся?

Netflix* утверждает, что мне точно понравится документальный фильм Bhutto, рассказывающий о жизни и трагической смерти бывшего пакистанского премьер-министра Беназир Бхутто. Возможно, мне действительно понравится этот фильм (я уже добавил его в список кинолент, которые собираюсь посмотреть). Прошлые рекомендации были просто потрясающими. К тому же когда Netflix советовала что-то из того, что я уже видел, то, как правило, фильм был из тех, которыми я действительно наслаждался.

Каким образом Netflix проделывает свои «фокусы»? Может быть, в штаб-квартире компании работает большое число стажеров, которые с помощью Google и опроса членов моей семьи и друзей «вычислили», что меня может заинтересовать документальный фильм о бывшем пакистанском премьер-министре? Конечно нет. Просто Netflix мастерски, со знанием дела использовала статистические данные. *Netflix не знакома со мной.* Но ей известно, какие фильмы мне понравились в прошлом (поскольку я выставлял им рейтинги). Воспользовавшись этой информацией наряду с рейтингами других кинозрителей и мощным компьютером, Netflix сумела сделать на удивление точные прогнозы относительно моих вкусов и предпочтений.

Я еще вернусь к алгоритму, который применила Netflix при составлении таких прогнозов, пока же достаточно будет сказать, что они основаны на корреляции. Netflix рекомендует фильмы, похожие на те, которые мне когда-то понравились или получили высокие оценки от других кинозрителей, чьи рейтинги подобны моим. Фильм Bhutto мне посоветовали потому, что в свое время я присвоил пятизвездочные рейтинги двум другим документальным фильмам: Enron: The Smartest Guys in the Room и Fog of War.

---

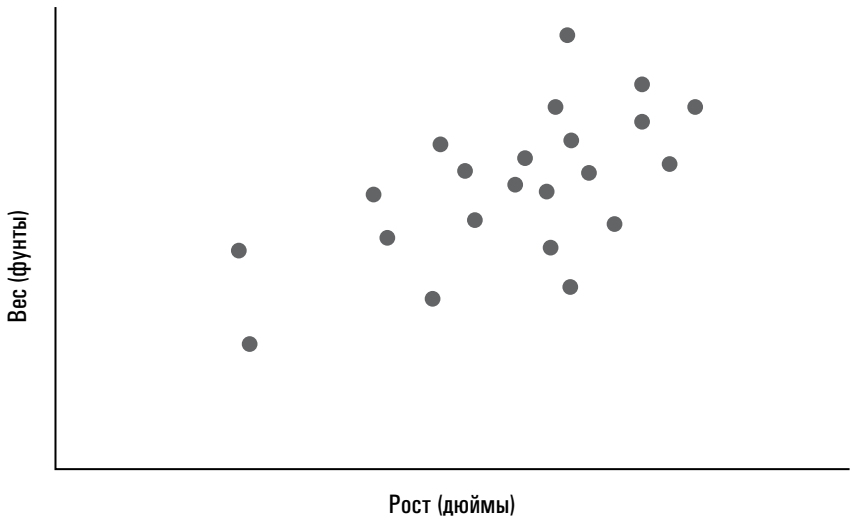
* Netflix — американская компания, поставщик фильмов и сериалов на основе потокового мультимедиа. *Прим. перев.*

Корреляция измеряет степень связи между двумя явлениями. Например, существует корреляция между летними температурами и продажей мороженого. Когда повышается температура, растут объемы продажи мороженого. Две переменные положительно коррелированы, если изменение одной переменной вызывает изменение другой в том же направлении, то есть в направлении увеличения или уменьшения (например, взаимосвязь между ростом и весом человека). У более высоких людей больший вес (в среднем); низкорослые люди весят меньше. Корреляция отрицательна, если положительное изменение одной переменной обусловливает отрицательное изменение другой (например, связь между регулярным выполнением физических упражнений и весом человека).

В зависимостях такого рода интересно то, что не каждое наблюдение вписывается в соответствующую схему. Иногда низкорослые люди весят больше, чем высокие. Иногда те, кто вообще не занимается спортом, бывают гораздо стройнее, чем те, кто регулярно выполняет физические упражнения. Тем не менее существует отчетливо выраженная связь между ростом и весом человека, а также между весом и физическими нагрузками.

Если построить диаграмму разброса данных, отражающих рост и вес произвольной выборки взрослых американцев, то получится примерно такая картина:

Диаграмма разброса данных, отражающих рост и вес человека



Если бы нам нужно было построить диаграмму разброса для данных о выполнении физических упражнений (количество минут, затрачиваемых на них каждую неделю) и данных о весе человека, то можно было бы ожидать отрицательной корреляции, причем те, кто занимается спортом больше времени, будут весить меньше. Однако картина в виде совокупности точек, разбросанных по определенной площади, представляет собой несколько неуклюжий инструмент. (Если бы Netflix попыталась предлагать мне какие-то фильмы, продемонстрировав диаграмму разброса рейтингов для тысяч кинолент, выставленных миллионами кинозрителей, то я посчитал бы такую рекомендацию просто неудачной шуткой.) Эффективность корреляции как статистического инструмента заключается в том, что мы можем выразить связь между двумя переменными с помощью одной описательной статистики — коэффициента корреляции.

Коэффициент корреляции обладает двумя чрезвычайно привлекательными характеристиками. Во-первых, в силу причин математического характера, которые мы обсудим в приложении, он представляет собой число в диапазоне от –1 до 1. Корреляция, равная 1 (иногда ее называют идеальной корреляцией), означает, что каждому изменению одной переменной соответствует эквивалентное изменение другой переменной в том же направлении.

Корреляция, равная –1 (иногда ее называют идеальной отрицательной корреляцией), означает, что каждому изменению одной переменной соответствует эквивалентное изменение другой переменной в противоположном направлении.

Чем ближе корреляция к 1 или –1, тем сильнее связь между переменными. Нулевая (или близкая к 0) корреляция говорит об отсутствии значимой связи между двумя переменными (например между результатом экзамена по математике и размером обуви экзаменуемого).

Второй привлекательной особенностью коэффициента корреляции является то, что с ним не связаны никакие единицы измерения. Мы можем рассчитать корреляцию между ростом и весом, несмотря на то что рост измеряется в дюймах, а вес — в фунтах. Мы можем даже вычислить корреляцию между количеством телевизоров, имеющихся дома у учеников, и результатами их экзаменов по математике (я почему-то уверен, что она окажется положительной). (Несколько ниже я остановлюсь подробнее на данной связи.) Коэффициент корреляции буквально творит чудеса: он сжимает сложное

сочетание данных, измеряемых в разных единицах (наподобие наших диаграмм разброса роста и веса), в единственную элегантную описательную статистику.

Как это удается?

Как обычно, я привожу самую распространенную формулу для определения коэффициента корреляции в приложении, находящемся в конце этой главы. Это не та статистика, которую можно вычислить вручную. (После того как вы введете соответствующие данные, базовый программный пакет, например Microsoft Excel, рассчитает корреляцию между двумя соответствующими переменными.) Тем не менее на интуитивном уровне понять эту формулу несложно. Формула для вычисления коэффициента корреляции выполняет следующие операции:

1. Вычисляет среднее значение и стандатное (среднеквадратическое) отклонение для обеих переменных. Если вернуться к примеру с ростом и весом, то мы бы узнали средний рост людей в выборке, средний вес людей в той же выборке и стандартное отклонение для роста и веса.
2. Преобразует все данные таким образом, чтобы каждое наблюдение было представлено его расстоянием (в стандартных отклонениях) от среднего значения. Верьте мне, это совсем не сложно. Допустим, средний рост в выборке равняется 66 дюймам (при стандартном отклонении в 5 дюймов), а средний вес — 177 фунтов (при стандартном отклонении в 10 фунтов). Теперь предположим, что ваш рост — 72 дюйма, а вес — 168 фунтов. Мы можем также сказать, что ваш рост составляет 1,2 стандартного отклонения сверх среднего роста [(72 - 66)/5) = 1,2] и 0,9 стандартного отклонения ниже среднего веса, или –0,9 применительно к нашей формуле [(168 - 177)/10 = –0,9]. *Да, это нетипично, когда рост человека выше среднего, а вес — ниже среднего, но поскольку вы уже заплатили неплохие деньги за эту книгу, то, как мне кажется, я должен в знак благодарности сделать вас высоким худощавым человеком.* Обратите внимание: ваш рост и вес, выражавшиеся поначалу в дюймах и фунтах, теперь выражаются абстрактными числами 1,2 и –0,9. Как видите, потребность в единицах измерения отпала.
3. Теперь я могу скрестить руки на груди и предоставить возможность компьютеру выполнить остальную работу. Формула вычисляет связь по всей выборке между ростом и весом, которые измеряются в стандартных единицах. Когда рост отдельных людей в выборке равняется,

к примеру, 1,5 или 2 стандартного отклонения выше среднего значения, какими должны быть значения их веса, *измеренные в стандартных отклонениях от среднего значения для веса?* А когда рост членов выборки близок к среднему значению, какими будут значения их веса, измеренные в стандартных единицах?

Если расстояние от среднего значения для одной переменной в целом соответствует — по величине и направлению — расстоянию от среднего значения для другой переменной (например, для людей, рост которых существенно отличается в ту или другую сторону от среднего значения роста, значения их веса, как правило, существенно отличаются от среднего значения веса, причем в том же направлении, что и рост), то у нас есть основания говорить о сильной положительной корреляции.

Если же расстояние от среднего значения для одной переменной в целом соответствует аналогичному расстоянию от среднего значения для другой переменной, *но в противоположном направлении* (например, у людей, которые чаще среднего занимаются физическими упражнениями, как правило, вес гораздо ниже среднего), то у нас есть основания говорить о сильной отрицательной корреляции.

Если две переменные в целом не отклоняются от среднего значения сколь-нибудь существенно (например, размер обуви и интенсивность занятий физическими упражнениями), то мы можем говорить о незначительной или нулевой корреляции.

Я чувствую, вы перенапряглись, читая этот раздел. Хочу вас утешить: вскоре мы вернемся к Netflix и тому, как ей удается угадывать ваш интерес к тем или иным фильмам. Однако вначале поразмышляем над еще одним событием, где корреляция играет немаловажную роль, — SAT. Да, именно SAT, о котором говорилось в главе 3. Этот тест (первоначальное название — Scholastic Aptitude Test) представляет собой стандартизированный экзамен, состоящий из трех разделов: математика, чтение и письмо. Возможно, вам уже приходилось его сдавать (или придется сдавать в будущем). Не исключено, что вы особо не задумывались над тем, *почему* вам нужно его сдавать. Цель этого экзамена — оценить вашу способность к обучению и спрогнозировать вашу успеваемость в колледже или университете. Разумеется, у вас (и особенно у тех из вас, кому не нравятся стандартизированные тесты) может возникнуть резонный вопрос: уж не для этого ли предназначена средняя школа? Почему так важен какой-то там четырехчасовой тест, если члены

приемной комиссии колледжа могли бы просто ознакомиться с оценками, которые вы получали на протяжении *четырех лет* учебы в старших классах школы?

Ответ на этот вопрос содержится в материале, с которым вы знакомились в главе 1 и 2. Оценки, которые выставляются ученикам в школе, представляют собой несовершенную описательную статистику. Ученик, получающий посредственные оценки при прохождении напряженной школьной программы для специализированных классов по математике и другим естественным наукам, может иметь бóльшие академические способности и потенциал, чем ученик той же школы, предпочевший программу с гуманитарным направлением. Это объясняется тем, что гуманитарные предметы усваиваются, как правило, гораздо легче, и получить высокие оценки по ним не составляет особого труда. Очевидно, что между разными школами также существуют немалые различия, которые сказываются на оценках учеников. Согласно данным College Board (орган, который разрабатывает и управляет SAT), этот тест призван «демократизировать доступ к высшим учебным заведениям для всех учащихся». Что можно возразить против такого довода? Все справедливо! SAT предлагает стандартизированный показатель способностей, который позволяет сравнивать всех абитуриентов, поступающих в колледжи и университеты. *Но можно ли считать его достаточно надежным показателем способностей?* Если мы хотим показатель, который позволяет легко сравнивать способности учащихся, то мы могли бы также предложить всем выпускникам школы посоревноваться в забеге на 100 ярдов, что было бы гораздо дешевле и проще, чем администрировать SAT. Проблема, конечно же, в том, что результат, показанный в забеге, никоим образом не коррелирован с академической успеваемостью в колледжах и университетах. Данные о результатах забега получить легко, однако они не имеют ничего общего с интересующим нас вопросом.

Чем же SAT лучше в этом отношении? К большому разочарованию будущих поколений старшеклассников, SAT вполне достойно справляется с задачей прогнозирования успехов студентов-первокурсников, так что сдавать его придется. College Board публикует соответствующие показатели корреляции. На шкале от 0 (полное отсутствие корреляции) до 1 (идеальная корреляция) корреляция между средней оценкой ученика старших классов школы и средней оценкой студента-первокурсника равняется 0,56. (Чтобы было понятнее, *что* это означает, скажу, что корреляция между ростом и весом

взрослых мужчин в Соединенных Штатах составляет примерно 0,4.) Корреляция между комплексным результатом, показанным при сдаче SAT (чтение, математика и письмо), и средним баллом студента-первокурсника также 0,56[1]. Это вроде бы говорит в пользу отказа от SAT, поскольку этот тест способен предсказать академическую успеваемость будущих студентов колледжей и университетов ничуть не лучше, чем средняя оценка ученика старших классов. По сути, самым надежным показателем будет комбинация баллов, полученных при сдаче SAT, и средней оценки ученика старших классов: корреляция между таким сочетанием и средним баллом студента-первокурсника составляет 0,64. Да, это действительно так.

Важным моментом в этом обсуждении является то, что корреляция не предполагает причинно-следственной связи: положительная или отрицательная корреляция между двумя переменными вовсе не обязательно означает, что изменения одной переменной вызывают изменения другой. Например, выше я указывал на вероятную положительную корреляцию между суммой баллов, полученных учащимся при сдаче SAT, и количеством телевизоров у него дома. Но это не значит, что родители могут существенно повысить результаты тестов своих детей путем покупки еще пяти телевизоров. Не говорит это, по-видимому, и о том, что сидение перед телевизором благотворно сказывается на академической успеваемости ученика.

Самым логичным объяснением такой корреляции может быть то, что высокообразованные родители могут себе позволить покупку нескольких телевизоров, что, однако, не мешает их детям сдавать экзамены с результатами, превышающими средний балл. Как количество телевизоров, так и экзаменационные оценки, по-видимому, обусловлены некой третьей переменной, коей является уровень образования родителей. Я не могу доказать наличие корреляции между количеством телевизоров в семье и количеством баллов, полученных при сдаче SAT (College Board не публикует соответствующих данных). Но готов доказать, что ученики из состоятельных семей демонстрируют в среднем более высокие результаты сдачи SAT, чем ученики из менее обеспеченных семей. Согласно данным, опубликованным College Board, учащиеся из семей с годовым доходом, превышающим 200 000 долларов, в среднем получают при сдаче математического раздела SAT 586 баллов, тогда как учащиеся из семей с годовым доходом, равным или меньшим 20 000 долларов, в среднем получают при сдаче того же математического раздела SAT лишь 460 баллов[2]. Между тем, вполне вероятно и то, что в домах

семей с годовым доходом, превышающим 200 000 долларов, больше телевизоров, чем в домах семей с годовым доходом менее 20 000 долларов.

Я начал писать эту главу несколько дней назад. За это время у меня появилась возможность посмотреть фильм Bhutto. Он действительно замечательный. Полная версия фильма, в которой охватывается период с момента отделения Пакистана от Индии в 1947 году до убийства пакистанского премьер-министра Беназир Бхутто в 2007-м, производит сильное впечатление. Голос Бхутто искусно вплетается в сюжетную линию в форме выступлений и интервью. Как бы то ни было, я пометил эту киноленту пятью звездочками, что вполне соответствует прогнозу Netflix.

В своей деятельности компания Netflix использует концепцию корреляции. Все началось с того, что я выставил оценки ряду фильмов. Netflix сравнила их с рейтингами других кинозрителей, чтобы выявить тех, чьи рейтинги высоко коррелированы с моими. Этим кинозрителям, как правило, нравятся те же фильмы, что и мне. Установив данный факт, Netflix может рекомендовать мне фильмы, которые понравились моим единомышленникам и которых я еще не видел.

Это, так сказать, «картина в целом». Фактическая методология гораздо сложнее. Вообще говоря, в 2006 году Netflix инициировала конкурс, в рамках которого обычным гражданам было предложено разработать механизм, который бы повысил эффективность уже существующих рекомендаций Netflix по меньшей мере на 10% (это означает, что данная система стала бы на 10% точнее при прогнозировании того, как бы кинозритель оценил тот или иной фильм после просмотра). Победителю был обещан 1 миллион долларов.

Каждый человек или группа людей, зарегистрировавшихся для участия в конкурсе, получал «обучающие данные», состоящие из более чем 100 миллионов рейтингов, выставленных 18 000 фильмам клиентами Netflix (их общее количество составляло 480 000 человек). Отдельная совокупность из 2,8 миллиона рейтингов не разглашалась (то есть Netflix знала, как кинозрители оценили эти фильмы, но участникам конкурса такая информация не предоставлялась). Конкурсантов оценивали по тому, насколько успешно предложенные ими алгоритмы прогнозировали фактические оценки, выставленные зрителями этих «неразглашенных» фильмов. Спустя три года тысячи команд из более чем 180 стран представили на суд жюри свои предложения. К участникам конкурса предъявлялось два требования. Во-первых, победитель должен был уступить Netflix права на свой алгоритм.

И во-вторых, он должен был «объяснить миру, как ему удалось решить эту задачу и каким образом она работает»[3].

В 2009 году Netflix объявила победителя. Им стала группа из семи человек, в состав которой входили статистики и программисты из США, Австрии, Канады и Израиля. Увы, я не могу описать здесь — даже в приложении — систему-победителя. Объяснение принципа ее действия занимает 92 страницы. Качество рекомендаций Netflix произвело на меня неизгладимое впечатление. Тем не менее система Netflix — просто супернавороченная вариация того, чем занимаются люди с момента появления кинематографа: найти кого-либо со схожими вкусами и попросить порекомендовать вам тот или иной фильм. Вам, как правило, нравятся те же фильмы, что и мне, и не нравятся те же фильмы, что и мне. Так что вы думаете о новом фильме Джорджа Клуни?

В этом и состоит суть корреляции.

Приложение к главе 4

Чтобы вычислить коэффициент корреляции между двумя совокупностями чисел, вы должны выполнить перечисленные ниже действия, каждое из которых иллюстрируется путем использования данных о значениях роста и веса для 15 гипотетических учащихся в приведенной ниже таблице.

1. Преобразуйте рост каждого учащегося в стандартные единицы: (рост – среднее значение) / стандартное отклонение.
2. Преобразуйте вес каждого из учащихся в стандартные единицы: (вес – среднее значение) / стандартное отклонение.
3. Для каждого учащегося вычислите произведение (вес в стандартных единицах) × (рост в стандартных единицах). Вы должны увидеть, что это число будет самым большим по абсолютному значению, когда рост *и* вес ученика расположены относительно далеко от своих средних значений.
4. Коэффициент корреляции представляет собой сумму произведений, вычисленных выше, деленную на количество наблюдений (в нашем случае — 15).

Корреляция между ростом и весом для этой группы учащихся — 0,83. Учитывая, что коэффициент корреляции может находиться в диапазоне от –1 до 1, это относительно высокая степень положительной корреляции, чего и следовало ожидать.

A *Учащийся*	*B* *Рост*	*C* *Вес*	*D* *Рост в стандартных единицах*	*E* *Вес в стандартных единицах*	*F* *(Вес в стандартных единицах) × (Рост в стандартных единицах)*
Ник	74	193	1,21	0,99	1,19
Элана	66	133	–0,63	–0,67	0,42
Дайна	68	155	–0,17	–0,06	0,01
Ребекка	69	147	0,06	–0,29	–0,02
Бен	73	175	0,98	0,49	0,48
Чару	70	128	0,29	–0,81	–0,24
Саха́р	60	100	–2,00	–1,59	3,18
Мэгги	63	128	–1,32	–0,81	1,07
Фейсал	67	170	–0,40	0,35	–0,14
Тед	70	182	0,29	0,68	0,20
Нарцисо	70	178	0,29	0,57	0,17
Катрина	70	118	0,29	–1,09	–0,32
Си Джей	75	227	1,44	1,93	2,77
София	62	115	–1,54	–1,17	1,81
Уилл	74	211	1,21	1,49	1,80
Среднее значение	68,73	157,33			Итого = 12,39
Стандартное отклонение	4,36	36,12		Коэффициент корреляции = Итого/n = 12,39/15 = 0,83	

Формула для вычисления коэффициента корреляции требует небольшого отступления, которое понадобится для того, чтобы объяснить систему обозначений, используемую в данном случае. Символ Σ часто применяется в статистике. Он обозначает суммирование величин, которые указаны после него. Если, например, имеется некая совокупность наблюдений x_1, x_2, x_3 и x_4, то запись $\Sigma\,(x_i)$ говорит о том, что мы должны суммировать четыре наблюдения: $x_1 + x_2 + x_3 + x_4$. Таким образом, $\Sigma\,(x_i) = x_1 + x_2 + x_3 + x_4$. Наша формула для среднего значения совокупности из n наблюдений может быть представлена в следующем виде: среднее значение = $\Sigma\,(x_i)/n$.

Мы можем придать этой формуле еще более универсальный вид, записав ее как $\sum_{i=1}^{n}(x_i)$. Эта формула означает суммирование величин $x_1 + x_2 + x_3 + \ldots + x_n$, или, другими словами, начиная с x_1 (поскольку $i = 1$) до x_n включительно

(поскольку $i = n$). Наша формула для среднего значения совокупности из n наблюдений может быть представлена в следующем виде:

среднее значение $= \sum_{i=1}^{n} (x_i)/n$.

С учетом этой универсальной системы обозначений формула вычисления коэффициента корреляции r для двух переменных x и y может выглядеть так:

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sigma_x} \frac{(y_i - \overline{y})}{\sigma_y},$$

где

n — количество наблюдений;

$\overline{x}_x$ — среднее значение для переменной x;

$\overline{y}_y$ — среднее значение для переменной y;

σ_x — стандартное отклонение для переменной x;

σ_y — стандартное отклонение для переменной y.

Любая статистическая компьютерная программа может с помощью статистических инструментов вычислить коэффициент корреляции между двумя переменными. Использование Microsoft Excel в примере с ростом и весом учащихся позволяет получить такую же корреляцию между ростом и весом пятнадцати учащихся, что и вычисление, выполненное нами вручную на основе приведенной выше таблицы: 0,83.

издательство
МАНН, ИВАНОВ И ФЕРБЕР
Максимально полезные книги