Условная вероятность. Теорема Байеса

В рассмотренных ранее примерах вычислялись вероятности элементарных событий. Возникает вопрос: как определить вероятность события, если известна некая информация о событиях, происшедших до него?[1] Вероятность события *A*, при вычислении которого учитывается информация о событии *B*, называется *условной* и обозначается как *P(A|B)*.

Вероятность события *A* при условии, что наступило событие *B*, равна вероятности события *A и B*, деленной на вероятность события *B*:

$$(1)\ P(A|B) = \frac{P(A \text{ и } B)}{P(B)}$$

Вероятность события *B* при условии, что наступило событие *A*, равна вероятности события *A и B*, деленной на вероятность события *A:*

$$(2)\ P(B|A) = \frac{P(A \text{ и } B)}{P(A)}$$

где *P(A и B)* – вероятность события *A и B*, *P(A)* – вероятность события *A*, *P(B)* – вероятность события *B*.

Фактически формулы (1) и (2) это краткая запись условной вероятности на основе таблицы сопряженности признаков. Вернемся к примеру, рассмотренному в предыдущей заметке (рис. 1). Предположим, что нам стало известно, будто некая семья собирается купить широкоэкранный телевизор. Какова вероятность того, что эта семья действительно купит такой телевизор?

	Совершена ли покупка?		
Планировалась ли покупка?	**Да**	**Нет**	**Всего**
Да	200	50	250
Нет	100	650	750
Всего	300	700	1 000

Рис. 1. Поведение покупателей широкоэкранных телевизоров

В данном случае нам необходимо вычислить условную вероятность Р (покупка совершена | покупка планировалась). Поскольку нам известно, что семья планирует покупку, выборочное пространство состоит не из всех 1000 семей, а только из тех, которые планируют покупку широкоэкранного телевизора. Из 250 таких семей 200 действительно купили этот телевизор. Следовательно, вероятность того, что семья действительно купит широкоэкранный телевизор, если она это запланировала, можно вычислить по следующей формуле:

Р (покупка совершена | покупка планировалась) = количество семей, планировавших и купивших широкоэкранный телевизор / количество семей, планировавших купить широкоэкранный телевизор = 200 / 250 = 0,8

Этот же результат дает формула (2):

$$P(B|A) = \frac{P(A \text{ и } B)}{P(A)}$$

где событие *A* заключается в том, что семья планирует покупку широкоформатного телевизора, а событие *B* — в том, что она его действительно купит. Подставляя в формулу реальные данные, получаем:

$$P\ (\text{покупка совершена} \mid \text{покупка планировалась}) = \frac{200\ /\ 1000}{250\ /\ 1000} = 0{,}8$$

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 265–279

Дерево решений

На рис. 1 семьи разделены на четыре категории: планировавшие покупку широкоэкранного телевизора и не планировавшие, а также купившие такой телевизор и не купившие. Аналогичную классификацию можно выполнить с помощью дерева решений (рис. 2). Дерево, изображенное на рис. 2, имеет две ветви, соответствующие семьям, которые планировали приобрести широкоэкранный телевизор, и семьям, которые не делали этого. Каждая из этих ветвей разделяется на две дополнительные ветви, соответствующие семьям, купившим и не купившим широкоэкранный телевизор. Вероятности, записанные на концах двух основных ветвей, являются безусловными вероятностями событий *A* и *A'*. Вероятности, записанные на концах четырех дополнительных ветвей, являются условными вероятностями каждой комбинации событий *A* и *B*. Условные вероятности вычисляются путем деления совместной вероятности событий на соответствующую безусловную вероятность каждого из них.

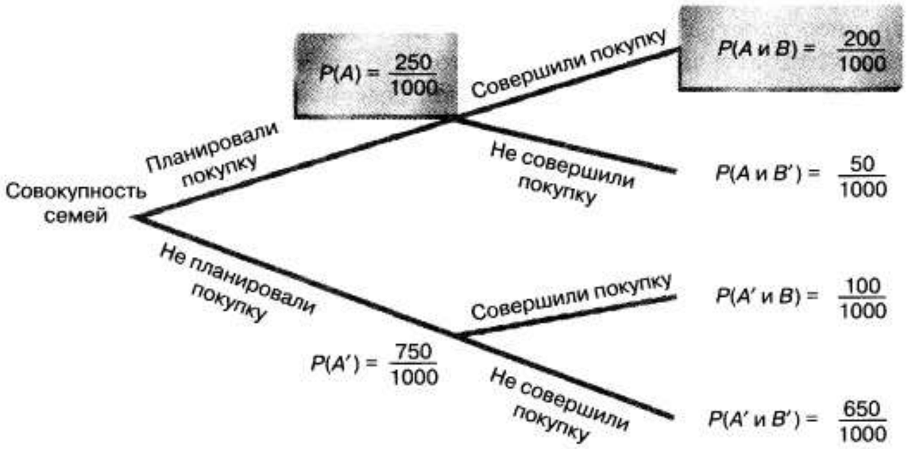

Рис. 2. Дерево решений

Например, чтобы вычислить вероятность того, что семья купит широкоэкранный телевизор, если она запланировала сделать это, следует определить вероятность события *покупка запланирована и совершена*, а затем поделить его на вероятность события *покупка запланирована*. Перемещаясь по дереву решения, изображенному на рис. 2, получаем следующий ответ:

$$\text{P (покупка совершена | покупка планировалась)} = \frac{200 / 1000}{250 / 1000} = 0{,}8$$

Статистическая независимость

В примере с покупкой широкоэкранного телевизора вероятность того, что случайно выбранная семья приобрела широкоэкранный телевизор при условии, что она планировала это сделать, равна 200/250 = 0,8. Напомним, что безусловная вероятность того, что случайно выбранная семья приобрела широкоэкранный телевизор, равна 300/1000 = 0,3. Отсюда следует очень важный вывод. *Априорная информация о том, что семья планировала покупку, влияет на вероятность самой покупки.* Иначе говоря, эти два события зависят друг от друга. В противоположность этому примеру, существуют статистически независимые события, вероятности которых не зависят друг от друга. Статистическая независимость выражается тождеством: $P(A|B) = P(A)$, где $P(A|B)$ — вероятность события *A* при условии, что произошло событие *B*, $P(A)$ — безусловная вероятность события А.

Обратите внимание на то, что события *A* и *B* являются статистически независимыми друг от друга тогда и только тогда, когда $P(A|B) = P(A)$. Если в таблице сопряженности признаков, имеющей размер 2x2, это условие выполняется хотя бы для одной комбинации событий *A* и *B*, оно будет справедливым и для любой другой комбинации. В нашем примере события *покупка запланирована* и *покупка совершена* не являются статистически независимыми, поскольку информация об одном событии влияет на вероятность другого.

Рассмотрим пример, в котором показано, как проверить статистическую независимость двух событий. Спросим у 300 семей, купивших широкоформатный телевизор, довольны ли они своей покупкой (рис. 3). Определите, связаны ли между собой степень удовлетворенности покупкой и тип телевизора.

Удовлетворены ли вы покупкой?			
Тип телевизора	Да	Нет	Всего
HDTV	64	16	80
Другой	176	44	220
Всего	240	60	300

Рис. 3. Данные, характеризующие степень удовлетворенности покупателей широкоэкранных телевизоров

Судя по этим данным,

$$\mathrm{P}\ (\text{покупатель удовлетворен} \mid \text{куплен HDTV-телевизор}\ ) = \frac{64\ /\ 300}{80\ /\ 300} = 0{,}80$$

В то же время,

$$\mathrm{P}\ (\text{покупатель удовлетворен}) = 240\ /\ 300 = 0{,}80$$

Следовательно, вероятность того, что покупатель удовлетворен покупкой, и того, что семья купила HDTV-телевизор, равны между собой, и эти события являются статистически независимыми, поскольку никак не связаны между собой.

Правило умножения вероятностей

Формула для вычисления условной вероятности позволяет определить вероятность совместного события *А и В*. Разрешив формулу (1)

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A\ и\ B)}}{\mathrm{P(B)}}$$

относительно совместной вероятности *Р(А и В)*, получаем общее, правило умножения вероятностей. Вероятность события *А и В* равна вероятности события *А* при условии, что наступило событие *В*, умноженной на вероятность события *В*:

$$(3)\ \mathrm{P(A\ и\ B) = P(A|B) * P(B)}$$

Рассмотрим в качестве примера 80 семей, купивших широкоэкранный HDTV-телевизор (рис. 3). В таблице указано, что 64 семьи удовлетворены покупкой и 16 — нет. Предположим, что среди них случайным образом выбираются две семьи. Определите вероятность, что оба покупателя окажутся довольными. Используя формулу (3), получаем:

$$\mathrm{P(A\ и\ B) = P(A|B) * P(B)}$$

где событие *А* заключается в том, что вторая семья удовлетворена своей покупкой, а событие *В* — в том, что первая семья удовлетворена своей покупкой. Вероятность того, что первая семья удовлетворена своей покупкой, равна 64/80. Однако вероятность того, что вторая семья также удовлетворена своей покупкой, зависит от ответа первой семьи. Если первая семья после опроса не возвращается в выборку (выбор без возвращения), количество респондентов снижается до 79. Если первая семья оказалась удовлетворенной своей покупкой, вероятность того, что вторая семья также будет довольна, равна 63/79, поскольку в выборке осталось только 63 семьи, удовлетворенные своим приобретением. Таким образом, подставляя в формулу (3) конкретные данные, получим следующий ответ:

Р(А и В) = (63/79)(64/80) = 0,638.

Следовательно, вероятность того, что обе семьи довольны своими покупками, равна 63,8%.

Предположим, что после опроса первая семья возвращается в выборку. Определите вероятность того, что обе семьи окажутся довольными своей покупкой. В этом случае вероятности того, что обе семьи удовлетворены своей покупкой одинаковы, и равны 64/80. Следовательно, Р(А и В) = (64/80)(64/80) = 0,64. Таким образом, вероятность того, что обе семьи довольны своими покупками, равна 64,0%. Этот пример показывает, что выбор второй семьи не зависит от выбора первой. Таким образом, заменяя в формуле (3) условную вероятность *P(A|B)* вероятностью *P(A)*, мы получаем формулу умножения вероятностей независимых событий.

Правило умножения вероятностей независимых событий. Если события *A* и *B* являются статистически независимыми, вероятность события *A и B* равна вероятности события *A*, умноженной на вероятность события *B*.

$$(4)\ P(A \text{ и } B) = P(A)P(B)$$

Если это правило выполняется для событий *A* и *B*, значит, они являются статистически независимыми. Таким образом, существуют два способа определить статистическую независимость двух событий:

1. События *A* и *B* являются статистически независимыми друг от друга тогда и только тогда, когда *P(A|B) = P(A)*.
2. События *A* и *B* являются статистически независимыми друг от друга тогда и только тогда, когда *P(A и B) = P(A)P(B)*.

Если в таблице сопряженности признаков, имеющей размер 2х2, одно из этих условий выполняется хотя бы для одной комбинации событий *A* и *B*, оно будет справедливым и для любой другой комбинации.

Безусловная вероятность элементарного события

$$(5)\ P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_k)P(B_k)$$

где события B_1, B_2, … B_k являются взаимоисключающими и исчерпывающими.

Проиллюстрируем применение этой формулы на примере рис.1. Используя формулу (5), получаем:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

где *P(A)* — вероятность того, что покупка планировалась, $P(B_1)$ — вероятность того, что покупка совершена, $P(B_2)$ — вероятность того, что покупка не совершена.

$$P(A) = \frac{200}{300}\frac{300}{1\,000} + \frac{50}{700}\frac{700}{1\,000} = \frac{200}{1\,000} + \frac{50}{1\,000} = \frac{250}{1\,000} = 0{,}25$$

ТЕОРЕМА БАЙЕСА

Условная вероятность события учитывает информацию о том, что произошло некое другое событие. Этот подход можно использовать как для уточнения вероятности с учетом вновь поступившей информации, так и для вычисления вероятности, что наблюдаемый эффект является следствием некоей конкретной причины. Процедура уточнения этих вероятностей называется теоремой Байеса. Впервые она была разработана Томасом Байесом в 18 веке.

Предположим, что компания, упомянутая выше, исследует рынок сбыта новой модели телевизора. В прошлом 40% телевизоров, созданных компанией, пользовались успехом, а 60% моделей признания не получили. Прежде чем объявить о выпуске новой модели, специалисты по маркетингу тщательно исследуют рынок и фиксируют спрос. В прошлом успех 80% моделей, получивших признание, прогнозировался заранее, в то же время 30% благоприятных прогнозов оказались неверными. Для новой модели отдел маркетинга дал благоприятный прогноз. Какова вероятность того, что новая модель телевизора будет пользоваться спросом?

Теорему Байеса можно вывести из определений условной вероятности (1) и (2). Чтобы вычислить вероятность P(B|A), возьмем формулу (2):

$$P(B|A) = \frac{P(A \text{ и } B)}{P(A)}$$

и подставим вместо P(A и B) значение из формулы (3):

$$P(A \text{ и } B) = P(A|B) * P(B)$$

Получим:

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

Подставляя вместо P(A) формулу (5), получаем теорему Байеса:

$$(6)\ P(B_i|A) = \frac{P(A|B_i) * P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)\ + \ldots + P(A|B_k)P(B_k)}$$

где события B_1, B_2, … B_k являются взаимоисключающими и исчерпывающими.

Введем следующие обозначения: событие S — *телевизор пользуется спросом*, событие S' — *телевизор не пользуется спросом*, событие F — *благоприятный прогноз*, событие F' — *неблагоприятный прогноз*. Допустим, что P(S) = 0,4, P(S') = 0,6, P(F|S) = 0,8, P(F|S') = 0,3. Применяя теорему Байеса получаем:

$$P(S|F) = \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|S')P(S')} = \frac{0{,}80 \times 0{,}40}{0{,}80 \times 0{,}40 + 0{,}30 \times 0{,}60} = \frac{0{,}32}{0{,}32 + 0{,}18} = \frac{0{,}32}{0{,}50} = 0{,}64$$

Вероятность спроса на новую модель телевизора при условии благоприятного прогноза равна 0,64. Таким образом, вероятность отсутствия спроса при условии благоприятного прогноза равна 1–0,64=0,36. Процесс вычислений представлен на рис. 4.

Событие S_i	***Априорная вероятность*** **$P(S_i)$**	***Условная вероятность*** **$P(F\|S_i)$**	***Совместная вероятность*** **$P(F\|S_i)\,P(S_i)$**	***Уточненная вероятность*** **$P(S_i\|F)$**
S — спрос есть	0,40	0,80	0,32	0,32/0,50 = 0,64 = $P(S\|F)$
S' — спроса нет	0,60	0,30	0,18	0,18/0,50 = 0,36 = $P(S'\|F)$
а			0,50	

б

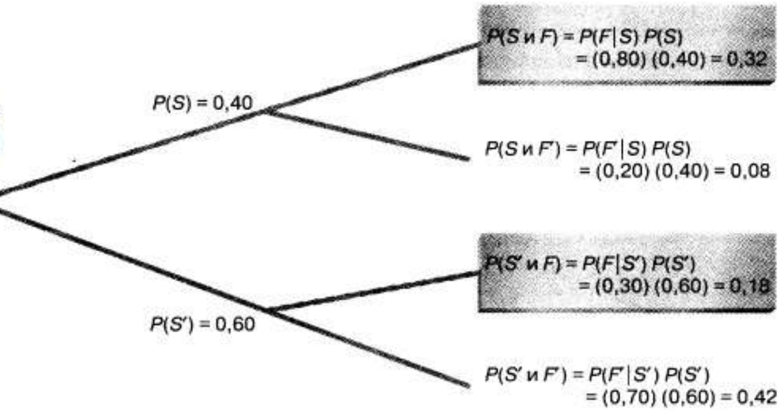

Рис. 4. (а) Вычисления по формуле Байеса для оценки вероятности спроса телевизоров; (б) Дерево решения при исследовании спроса на новую модель телевизора

Рассмотрим пример применения теоремы Байеса для медицинской диагностики. Вероятность того, что человек страдает от определенного заболевания, равна 0,03. Медицинский тест позволяет проверить, так ли это. Если человек действительно болен, вероятность точного диагноза (утверждающего, что человек болен, когда он действительно болен) равна 0,9. Если человек здоров, вероятность ложноположительного диагноза (утверждающего, что человек болен, когда он здоров) равна 0,02. Допустим, что медицинский тест дал положительный результат. Какова вероятность того, что человек действительно болен? Какова вероятность точного диагноза?

Введем следующие обозначения: событие D — *человек болен*, событие D' — *человек здоров*, событие T — *диагноз положительный*, событие T' — *диагноз отрицательный*. Из условия задачи следует, что P(D) = 0,03, P(D') = 0,97, P(T|D) = 0,90, P(T|D') = 0,02. Применяя формулу (6), получаем:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')} = \frac{0{,}90 \times 0{,}03}{0{,}90 \times 0{,}03 + 0{,}02 \times 0{,}97} = \frac{0{,}0270}{0{,}0464} = 0{,}582$$

Вероятность того, что при положительном диагнозе человек действительно болен, равна 0,582 (см. также рис. 5). Обратите внимание на то, что знаменатель формулы Байеса равен вероятности положительного диагноза, т.е. 0,0464.

Событие D_i	*Априорная вероятность* $P(D_i)$	*Условная вероятность* $P(T\|D_i)$	*Совместная вероятность* $P(T\|D_i)P(D_i)$	*Уточненная вероятность* $P(D_i\|T)$
D — болен	0,03	0,90	0,0270	$0,0270/0,0464 = 0,582 = P(D\|T)$
D' — здоров	0,97	0,02	0,0194	$0,0194/0,0464 = 0,418 = P(D'\|T)$
а			0,0494	

б

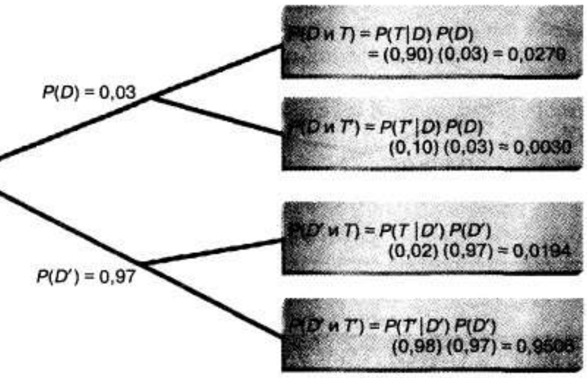

Рис. 5. (а) Вычисления по формуле Байеса для оценки точности медицинского диагноза; (б) Дерево решения при оценке точности медицинского диагноза

Предыдущая заметка [Основные понятия теории вероятностей]

Следующая заметка

К оглавлению [Статистика для менеджеров с использованием Microsoft Excel]

Возможно, вас также заинтересует:

[Дуглас Хаббард. Как измерить всё, что угодно. Оценка стоимости нематериального в бизнесе]

[Леонард Млодинов. (Не)совершенная случайность. Как случай управляет нашей жизнью]

[Канеман, Словик, Тверски. Принятие решений в неопределенности: Правила и предубеждения]