1. Правило Ассоциации: Ex. {X → Y} - это представление нахождения Y в корзине, в которой есть X

Транзакции – это наборы последовательных действий

**Support** указание того, как часто элемент появляется в наших данных в транзакциях

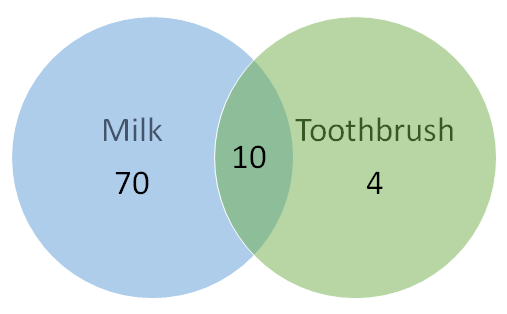
Ценность поддержки помогает нам определить правила, которые стоит рассмотреть для дальнейшего анализа. Например, можно было бы рассмотреть только наборы элементов, которые происходят по крайней мере 50 раз из общего числа 10 000 транзакций, т. е. поддержка = 0.005. Если набор элементов имеет очень низкую поддержку, мы не располагаем достаточной информацией о взаимосвязи между его элементами и, следовательно, не можем сделать никаких выводов из такого правила

**Confidence** количество правил if then которые были true

С технической точки зрения уверенность-это условная вероятность возникновения консеквента, заданного антецедентом.

Давайте рассмотрим еще несколько примеров, прежде чем двигаться дальше. Как вы думаете, какова была бы уверенность для {масло} → {хлеб}? То есть, какая доля сделок, имеющих масло, также имела хлеб? Очень высокое, то есть значение, близкое к 1? Вот именно. А как насчет {йогурт} → {молоко}? Опять под кайфом. {Зубная Щетка} → {Молоко}? Не совсем уверен? Доверие к этому правилу также будет высоким, поскольку {молоко} является таким частым набором элементов и будет присутствовать в каждой другой транзакции.

*Не имеет значения, что вы имеете в предшествующем для такого частого следствия. Доверие к ассоциативному правилу, имеющему очень частое следствие, всегда будет высоким*



Рассмотрим цифры из рисунка слева. Уверенность для {зубной щетки} → {молока} будет равна 10/(10+4) = 0,7

Похоже на высокое значение доверия. Но мы интуитивно знаем, что эти два продукта имеют слабую ассоциацию, и есть что-то обманчивое в этом высоком доверительном значении. Лифт вводится для преодоления этой проблемы.

If если купил хлеб ты купишь молоко

Lift сколько раз if then будет предположительно истинным (сравнение конфиденс с ожидаемым)

Думайте об этом как о "подъемнике", который {X} обеспечивает нашу уверенность в том, что {Y} находится на тележке. Перефразируя, подъем-это повышение вероятности наличия {Y} на тележке со знанием присутствия {X} над вероятностью наличия {Y} на тележке без какого-либо знания о присутствии {X}

В тех случаях, когда {X} фактически ведет к {Y} на тележке, значение подъема будет больше 1. Давайте разберемся в этом на примере, который будет продолжением правила {зубная щетка} → {молоко}.

Вероятность наличия молока на тележке при знании того, что зубная щетка присутствует (т. е. уверенность) : 10/(10+4) = 0.7

Теперь, чтобы представить это число в перспективе, рассмотрим вероятность наличия молока на тележке без каких-либо знаний о зубной щетке: 80/100 = 0,8

Эти цифры показывают, что наличие зубной щетки на тележке фактически снижает вероятность наличия молока на тележке до 0,7 с 0,8! Это будет подъем на 0,7 / 0,8 = 0,87. Теперь это больше похоже на реальную картину. Значение подъема менее 1 показывает, что наличие зубной щетки на тележке не увеличивает шансов появления молока на тележке, несмотря на правило, показывающее высокое доверительное значение. Значение подъема больше 1 гарантирует высокую ассоциацию между {Y} и {X}. Чем больше значение подъема, тем больше шансов предпочтения купить {Y}, если клиент уже купил {X}.

Apriori – для предварительного знания о частых свойствах набора элементов, то есть алгоритм говорит есть построенные правила

создание правил из всего списка элементов и определение наиболее важных из них

40 – Если confidence = 0 выводит все правила, а саппорт = 0,01 частота

64 - Чем выше конфиденс тем меньше правмил будет сгенерировано

Чем выше конфиденс и лифт тем интереснее правило тем сильнее правило (то есть вероятность того, что он выполнится, это лифт, а конфиленс это

85 – Тут четко видны что слева пустые правила

114 – Самое лучшее правило самое темное, синее не является самым лучшим, не смотря на конфиденс, а их много, можно отбросить

132 – Првила которые левые от итем до круглишка, а справа то что от круглишка до итемма

Последнее является ли наш алгроитм

количество случайных наборов, если центры - целое число. Проверяются случайные центры nstart, и возвращается наилучшее соответствие (с точки зрения минимальной суммы в пределах суммы квадратов кластера).

В иерархической лкастеризации ближайшие две точки, потом берет 3 и считает ее как одну

Берется небольшая выборкаь

Тестовый набор данных проверяю свои данные, нормальное ли у нас разбиение или данные

Тренировочный рассматриваю модели, априори

ты берешь 80% от датасета. тренируешь модель, то есть выставляешь нужные теперь параметры в функциях и алгоритмах. прогоняешь и получаешь результат. потом прогоняешь эту же модель (с теми же параметрами) на тестовой модели. сравнивашеь результаты. если они совпадают (типо распределение и тд), то модель обучена норм

* ты должен сравнивать не количетсво

а процент



то есть и там и там правил под 1000 должно быть



если совпадает 22, то хуево



у меня совпадало 50%



и тоже плохо