5. ЛЕКЦИЯ. Интеллектуальный анализ данных: байесовский классификатор.

Простой байесовский классификатор

Байесовский подход представляет собой группу алгоритмов классификации, основанных на принципе максимума апостериорной вероятности. Сначала определяется апостериорная вероятность (условная вероятность случайного события при условии того, что известны апостериорные данные, т.е. полученные после опыта.) отношения объекта к каждому из классов, а затем выбирается тот класс, для которого она максимальна. Алгоритм предсказывает вероятность того, что объект или наблюдение относится к определенному классу.

Этот подход к классификации является одним из старейших и до сих пор сохраняет прочные позиции в технологиях анализа данных. Кроме того, он лежит в основе многих удачных алгоритмов классификации. Рассмотрим один из них так называемый простой, или наивный, байесовский классификатор. Это специальный случай байесовского классификатора, в котором используется предположение о статистической независимости признаков описывающих классифицируемые объекты. Такое предположение существенно упрощает задачу, поскольку вместо одной многомерной плотности вероятности по всем признакам достаточно оценить несколько одномерных плотностей. К сожалению, на практике предположение о независимости признаков редко выполняется, является «наивным», что и дало название методу.

Основные преимущества наивного байесовского классификатора: легкость программной реализации и низкие вычислительные затраты при обучении и классификации. В тех редких случаях, когда признаки действительно независимы (или близки к этому), он почти оптимален. Главный его недостаток относительно низкое качество классификации в большинстве реальных задач. Поэтому чаще всего его используют либо как примитивный эталон для сравнения различных моделей, либо как блок для построения более сложных алгоритмов.

Рассмотрим базовые принципы работы простого байесовского классификатора.

Основные теоретические сведения

Пусть имеется объект или наблюдение X, класс которого неизвестен. Пусть также имеется гипотеза H, согласно которой X относится к некоторому классу C. Для за дачи классификации можно определить вероятность $P(H|X)$, то есть вероятность того, что гипотеза H для X справедлива. $P(H|X)$ называется условной вероятностью того, что гипотеза H верна при условии, что классифицируется объект X, или апостериорной вероятностью.

(| - множество элементов, удовлетворяющих условию, множество всех… таких, что верно…). Апостериорное распределение – условное распределение вероятностей какой-либо случайной величины при некотором условии, рассматриваемое в противоположность ее безусловному или априорному распределению)

Предположим, что объектами классификации являются фрукты, которые описываются их цветом и размером, Определим объект X как красный и круглый и выдвинем гипотезу H, что это яблоко. Тогда условная вероятность $P(H|X)$ отражает меру уверенности в том, что объект X является яблоком при условии, что он красный и круглый, Кроме условной (апостериорной) вероятности, рассмотрим так называемую априорную вероятность $P(H)$. В нашем примере это вероятность того, что любой наблюдаемый объект является яблоком, безотносительно к тому, как он выглядит. Таким образом, апостериорная вероятность основана на большей информации, чем априорная, не предполагающая зависимость от свойств объекта X.

Аналогично $P(H|X)$ апостериорная вероятность X при условии H, или вероятность того, что X является красным и круглым, если известно, что это яблоко, $P(X)$априорная вероятность X. В нашем примере это просто вероятность того, что объект является красным и круглым. Вероятности $P(X)$, $P(H)$ и $P(X|H)$ могут быть оценены на основе наблюдаемых данных.

Для вычисления апостериорной вероятности на основе $P(X)$, $P(H)$ и $P(X|H)$ используется формула Байеса:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Алгоритм работы простого байесовского классификатора содержит следующие шаги.

1. Пусть исходное множество данных S содержит атрибуты A_1, A_2, ..., A_n Тогда каждый объект или наблюдение $X \in S$ будет представлено своим набором значений этих атрибутов $x_1, x_2, \ldots, x_n$ где x_i значение, которое принимает атрибут A_i в данном наблюдении.

2. Предположим, что задано m классов $C = \{C_1, C_2, \ldots C_m\}$ и наблюдение X, для которого класс неизвестен. Классификатор должен определить, что X относится к классу, который имеет наибольшую апостериорную вероятность $P(H|X)$. Простой байесовский классификатор относит наблюдение X к классу $C_k(k = 1, \ldots, m)$ тогда и только тогда, когда выполняется условие $P(C_k|X) > P(C_j|X)$ для любых $1 \le j \le m$:.т.к, $k \ne j$.

По формуле Байеса:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad (2)$$

3. Поскольку вероятность $P(X)$ для всех классов одинакова, максимизировать требуется только числитель формулы (2), Если априорная вероятность класса $P(C_k)$ неизвестна, то можно предположить, что классы равновероятны, $P(C_1) = P(C_2) = \ldots = P(C_m)$, и, следовательно, мы должны выбрать максимальную вероятность $P(C_k|X)$.

Заметим, что априорные вероятности классов могут быть оценены как $P(C_k) = s_k/s$, где s_k – число наблюдений обучающей выборки, которые относятся к классу C_k, а s – общее число обучающих примеров.

4. Если исходное множество данных содержит большое количество атрибутов, то определение $P(X|C_k)$ может потребовать значительных вычислительных затрат. Чтобы их уменьшить, используется «наивное» предположение о независимости признаков. То есть для набора атрибутов $X = (x_1, x_2, \ldots, x_n)$ можно записать:

$$P(X|C_k) = P(x_1|C_k) \times P(x_2|C_k) \times \ldots \times P(x_n|C_k) \quad (3)$$

Вероятности, стоящие в правой части формулы (3), могут быть определены из обучающего набора данных для следующих случаев.

Атрибут A является категориальным, тогда $P(X|C_k) = {s_{ik}}/{s_k}$, где s_{ik} – общее число наблюдений класса C_i, в которых A_i принимает значение x_i, а s_k – общее число наблюдении, относящихся к классу C_k.

Атрибут A является непрерывным, тогда предполагается, что его значения подчиняются закону распределения Гаусса:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(\frac{(x-m)^2}{2\sigma^2}\right)$$

где m и σ^2 – математическое ожидание и дисперсия значений атрибута A_i для наблюдений, относящихся к классу C_k.

При классификации неизвестного наблюдения объект X будет относиться к классу, для которого $P(X|C_i) \times P(C_i)$ принимает наибольшее значение.

Пример работы простого байесовского классификатора

Рассмотрим пример работы наивного байесовского классификатора на основе задачи предсказания текучести абонентской базы, в качестве набора данных 14 записей с квантованными входными полями (табл. 1).

Таблица 1. Набор данных «Текучесть абонентской базы»

№	Количество звонков	SMS-активность	Internet-активность	Международные звонки	Уход

1	< 100	Высокая	Средняя	Нет	Да
2	< 100	Высокая	Высокая	Нет	Да
3	$100 - 150$	Высокая	Средняя	Нет	Нет
4	> 150	Средняя	Средняя	Нет	Да
5	> 150	Низкая	Средняя	Да	Нет
6	> 150	Низкая	Высокая	Да	Нет
7	$100 - 150$	Низкая	Высокая	Да	Да
8	< 100	Средняя	Средняя	Нет	Нет
9	< 100	Низкая	Средняя	Да	Да
10	> 150	Средняя	Средняя	Да	Да
11	< 100	Средняя	Высокая	Да	Да
12	$100 - 150$	Средняя	Высокая	Нет	Да
13	$100 - 150$	Высокая	Средняя	Да	Да
14	> 150	Средняя	Высокая	Нет	Нет

Обозначим как C_1 класс клиентов, для которых Уход = Да. К классу C_2 будем относить клиентов, которые сохраняют лояльность компании. Для них метка класса будет Нет. Пусть новый клиент, вероятность ухода которого мы хотим оценить, обладает следующими параметрами:

$$(\text{X} = (\text{Количество звонков} =< 100, SMS - \text{активность} = \text{Средняя}, Internet - \text{активность} = \text{Средняя}, \text{Международные звонки} = \text{Да}).$$

Согласно простому алгоритму Байеса нужно максимизировать $P(X|C_k) \times P(C_k)$ для k = 1,2 (поскольку классов всего два). Априорная вероятность появления каждого $P(C_k)$ может быть вычислена с помощью обучающего множества из таблицы как отношение числа примеров -го класса к общему числу примеров. Поскольку всего примеров 14, наблюдений с меткой класса Да - 9, а с меткой Нет - 5, то $P(C_2) = 5/14 = 0{,}3\,57.$

Для вычисления $P(X|C_k)$, $k = 1{,}2$ определим следующие условные вероятности (табл.2.).

Таблица 2. - Расчет условных вероятностей

$\text{P}(\text{Число звонков} =< 100|\text{Уход} = \text{Да}) = 4/9 = 0{,}444$
$\text{P}(\text{Число звонков} =< 100|\text{Уход} = \text{Нет}) = 1/5 = 0{,}2$
$\text{P}(SMS - \text{активность} = \text{Средняя}|\text{Уход} = \text{Да}) = 4/9 = 0{,}444$
$\text{P}(SMS - \text{активность} = \text{Средняя}|\text{Уход} = \text{Нет}) = 2/5 = 0{,}4$
$\text{P}(Internet - \text{активность} = \text{Средняя}|\text{Уход} = \text{Да}) = 5/9 == 0{,}555$
$\text{P}(Internet - \text{активность} = \text{Средняя}|\text{Уход} = \text{Нет}) = 3/5 == 0{,}6$

$\text{P(Международные звонки} = \text{Да}|\text{Уход} = \text{Да}) = 5/9 = 0{,}555$
$\text{P(Международные звонки} = \text{Да}|\text{Уход} = \text{Нет}) = 2/5 = 0{,}4$

Используя рассчитанные условные вероятности, получим (табл. 3).

Таблица 3. Расчет условных вероятностей

$\text{P(X}|\text{Уход} = \text{Да}) = 0{,}444 \times 0{,}444 \times 0{,}555 \times 0{,}555 = 0{,}061$
$\text{P(X}|\text{Уход} = \text{Нет}) = 0{,}6 \times 0{,}4 \times 0{,}2 \times 0{,}4 = 0{,}019$
$\text{P(X}|\text{Уход} = \text{Да}) \times \text{P(Уход} = \text{Да}) = 0{,}061 \times 0{,}643 = 0{,}039$
$\text{P(X}|\text{Уход} = \text{Нет}) \times \text{P(Уход} = \text{Нет}) = 0{,}019 \times 0{,}357 = 0{,}007$

Таким образом, для наблюдения X произведение вероятностей для класса C_1, то есть $P(X|C_1) \times P(C_1) = 0{,}039$, а для класса C_2 $P(X|C_1) \times P(C_1) = 0{,}007$, Поэтому выбирается класс C_1, для которого оно больше. Можно нормализовать эти вероятности, тогда получим: $P/(X|C_1) \times P(C_1) = 0{,}039/(0{,}039 + 0{,}007) = 0{,}848$ и $P/(X|C_2) \times P(C_2) = 0{,}007/(0{,}039 + 0{,}007) = 0{,}152$

Следовательно, для наблюдения X метка класса будет Да, а клиент, характеризуемый соответствующими признаками, должен рассматриваться как склонный к уходу.

Байесовские сети

Проблема заключается в том, что распределения, которые нас интересуют, обычно слишком сложные, чтобы их можно было максимизировать напрямую, аналитически. В них слишком много переменных, между переменными слишком сложные связи. Но, с другой стороны, часто в них есть дополнительная структура, которую можно использовать, структура в виде независимостей $(p(x, y) = p(x)p(y))$ и условных независимостей $(p(x, y|\,\text{z}) = p(x|\text{z})p(y|z))$ некоторых переменных

В результате сложное апостериорное распределение $(p(a_1, \dots, a_n|\, x = v)p(x = v)$ удалось переписать как $p(x = v)p(a_1|x = v)p(a_2|x = v) \dots p(a_n|x = v)$ и в этой форме с ним гораздо проще справиться (обучить параметры каждого маленького распределения по отдельности, а затем выбрать v, дающее максимум произведения. Это один из простейших примеров разложения (факторизации) сложного распределения в произведение простых.

Итак, давайте рассмотрим один из самых удобных способов представлять большие и сложные распределения вероятностей – байесовские сети доверия, которые в последнее время чаще называются просто направленными графическими моделями.

Байесовская сеть – это направленный граф без направленных циклов (*это очень важное условие!*), в котором вершины соответствуют переменным в

распределении, а рёбра соединяют «связанные» переменные. В каждом узле задано условное распределение узла при условии своих родителей $p(x|parents(x))$ (родители), и граф байесовской сети означает, что большое совместное распределение раскладывается в произведение этих условных распределений. Вот, например, граф, соответствующий наивному Байесу (рис.1):

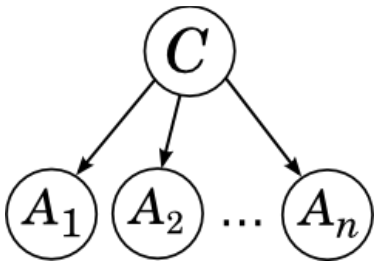


Рис.1. Граф, соответствующий наивному Байесу

Он соответствует разложению $p(A_1, \dots, A_n, C) = p(C)p(A_1|C)p(A_2|C) \dots p(A_n|C))$: у C нет предков, так что мы берём его безусловное распределение, а каждый из A_i «растёт» непосредственно из C и больше ни с кем не связан.

Мы уже знаем, что в этом случае все атрибуты A_i условно независимы при условии категории C:

$$p(A_i, A_j|C) = p(A_i|C)p(A_j|C)$$

Давайте теперь рассмотрим все простейшие варианты байесовской сети и посмотрим, каким условиям независимости между переменными они соответствуют.

Байесовские сети с двумя и тремя переменными: тривиальные (упрощенные) случаи

Начнём с сети из двух переменных, x и y. Здесь всего два варианта: либо между x и y нет ребра, либо есть. Если ребра нет, это просто значит, что x и y независимы (рис.2), ведь такой граф соответствует разложению $p(x, y) = p(x)p(y)$

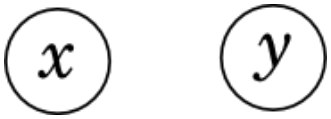


Рис.2. Переменные x и y независимы

А если ребро есть (пусть оно идёт из x в y, это не важно), мы получаем разложение $p(x, y) = p(x)p(y|x)$, которое буквально по определению условной вероятности тупо верно всегда, для любого распределения $p(x, y)$. Таким образом, граф из двух вершин с ребром не даёт нам новой информации (рис.3).

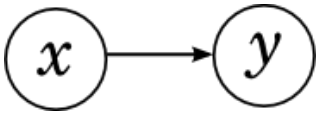


Рис.2. Переменные x и y зависимы

Теперь переходим к сетям из трёх переменных x, y и z. Самый простой случай – когда рёбер совсем нет.

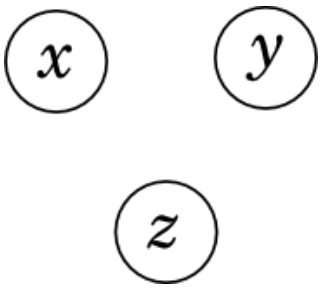


Рис.3. Переменные x, y и z независимы

Как и с двумя переменными, это значит, что x, y и z просто независимы: $p(x, y, z) = p(x)p(y)p(z)$.Другой простой случай – когда между переменными проведены все рёбра (рис.4).

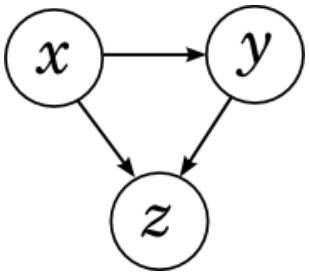


Рис.4. Переменные x, y и z независимы

Этот случай тоже аналогичен рассмотренному выше; пусть, например, рёбра идут из x в y и z, а также из y в z. Получаем разложение $p(x, y, z) = p(x)p(y|x)p(z|\mathrm{x}, \mathrm{y})$ которое, опять же, верно всегда, для любого совместного распределения $p(x, y, z)$. В этой формуле можно было бы выбирать переменные в любом порядке, ничего не изменилось бы. Обратите внимание, что направленные циклы в байесовских сетях запрещены, и в результате вариантов, как можно провести все рёбра, всего шесть, а не восемь.

Рассмотрим три более интересных случая – это и будут те «кирпичики», из которых можно составить любую байесовскую сеть. К счастью, для этого достаточно рассмотреть графы на трёх переменных – всё остальное будет обобщаться из них.

В примерах ниже будут интуитивно интерпретировать ребро, стрелочку между двумя переменными, как «x влияет на y», т.е. по сути, как причинно-следственную связь. На самом деле это, конечно, не совсем так.

Последовательная связь

Начнём с последовательной связи между переменными: x «влияет на» y, а y, в свою очередь, «влияет на» z (рис.5).

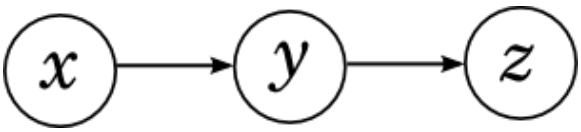


Рис.5. Последовательная связь переменных x, y и z

Такой граф изображает разложение $p(x, y, z) = p(x)p(y|x)p(z|y)$

Интуитивно это соответствует последовательной причинно-следственной связи: если вы будете бегать зимой без шапки, вы простудитесь, а если простудитесь, у вас поднимется температура. Очевидно, что x и y, а также y и z друг с другом связаны, между ними даже непосредственно стрелочки проведены. Связаны ли между собой в такой сети x и z, зависимы ли эти переменные? Конечно! Если вы бегаете зимой без шапки, вероятность получить высокую температуру повышается. Однако в такой сети x и z связаны только через y, и если мы уже знаем значение y, x и z становятся независимыми: если вы уже знаете, что простудились, совершенно не важно, чем это было вызвано, температура теперь повысится (или не повысится) именно от простуды.

Формально это соответствует условной независимости x и z при условии y; давайте это проверим:

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

где первое равенство – это определение условной вероятности, второе – наше разложение, а третье – применение теоремы Байеса.

Итак, последовательная связь между тремя переменными говорит нам о том, что крайние переменные условно независимы при условии средней. Всё очень логично и достаточно прямолинейно.

Расходящаяся связь

Следующий возможный вариант – расходящаяся связь: x «влияет» и на y, и на z. (рис.6)

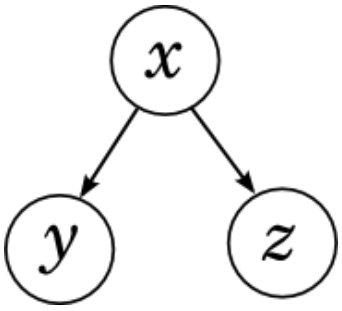


Рис.6. Расходящаяся связь переменных x, y и z

Такой граф изображает разложение

$$p(x, y, z) = p(x)p(y|x)p(z|x)$$

Интуитивно это соответствует двум следствиям из одной и той же причины: если вы простудитесь, у вас может подняться температура, а также может начаться насморк. Как и в предыдущем случае, очевидно, что x и y, а также x и z зависимы, и вопрос заключается в зависимости между y и z. Опять же, очевидно, что эти переменные зависимы: если у вас насморк, это повышает вероятность того, что вы простудились, а значит, вероятность высокой температуры тоже повышается.

Однако в такой сети, подобно предыдущему случаю, y и z связаны только через x, и если мы уже знаем значение общей причины x, y и z становятся независимыми: если вы уже знаете, что простудились, насморк и температура становятся независимы.

Формально это соответствует условной независимости y и z при условии x; проверить это ещё проще, чем для последовательной связи:

$$p(y,z|x) = \frac{p(x,y,z)}{p(y)} = \frac{p(x)p(y|x)p(z|\mathrm{x})}{p(y)} = p(y|x)p(z|x)$$

Итак, расходящаяся связь между тремя переменными говорит нам о том, что «следствия» условно независимы при условии своей «общей причины». Если причина известна, то следствия становятся независимы; пока причина неизвестна, следствия через неё связаны.

Сходящаяся связь

У нас остался только один возможный вариант связи между тремя переменными: сходящаяся связь, когда x и y вместе «влияют на» z.

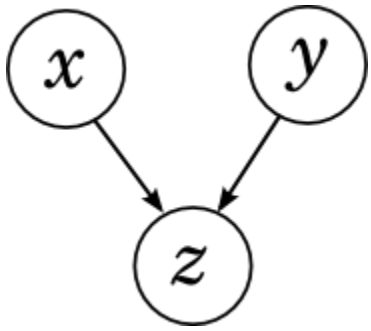


Рис.6. Сходящаяся связь переменных x,y и z

Разложение здесь получается такое:

$$p(x,y,z) = p(x)p(y)p(z|x,y)$$

Это ситуация, в которой у одного и того же следствия могут быть две разные причины: например, температура может быть следствием простуды, а может – отравления. Зависимы простуда и отравление? Нет! В этой ситуации, пока общее следствие неизвестно, две причины никак не связаны друг с другом, и это очень легко проверить формально:

$$p(x,y) = \sum_z p(x,y,z) = \sum_z p(x), p(y), p(z|x,y) = p(x)p(y)$$

Однако если «общее следствие» z становится известным, ситуация меняется. Теперь общие причины известного следствия начинают влиять друг на друга. Предположим, что вы знаете, что у вас температура. Это сильно повышает вероятность, как простуды, так и отравления. Однако если вы теперь узнаете, что отравились, вероятность простуды уменьшится – симптом «уже объяснён» одной из возможных причин, и вторая становится менее вероятной.

Таким образом, при сходящейся связи две «причины» независимы, но только до тех пор, пока значение их «общего следствия» неизвестно; если же общее следствие получает означивание, причины становятся зависимыми.

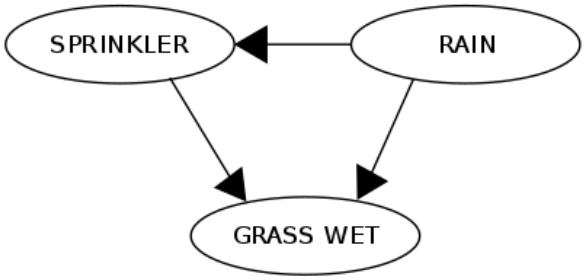
Но есть один нюанс: когда на пути встречается сходящаяся связь, недостаточно посмотреть только на её «общее следствие», чтобы определить независимость. На самом деле, если даже у z означивания нету, но оно есть у одного из её потомков (возможно, достаточно далёких), две причины всё равно станут зависимы. Интуитивно это тоже легко понять: например, пусть мы не наблюдаем собственно температуру, а наблюдаем её потомка – показания градусника. На свете, наверное, бывают неисправные градусники, так что это тоже некая вероятностная связь. Однако наблюдение показаний градусника точно так же делает простуду и отравление зависимыми.

Последовательная, сходящаяся и расходящаяся связи – это те три кирпичика, из которых состоит любой ациклический направленный граф. И наших рассуждений вполне достаточно для того, чтобы обобщить результаты об условной зависимости и независимости на все такие графы. Конечно, здесь не время и не место для того, чтобы формально доказывать общие теоремы, но результат достаточно предсказуем – предположим, что вам нужно в большом графе проверить, независимы ли две вершины (или даже два множества вершин). Для этого вы смотрите на все пути в графе (без учёта стрелочек), соединяющие эти два множества вершин. Каждый из этих путей можно «разорвать» одной из вышеописанных конструкций: например, последовательная связь разорвётся, если в середине у неё есть означивание (значение переменной известно), а сходящаяся связь разорвётся, если, наоборот, означивания нет (причём нет ни в самой вершине из пути, ни в её потомках). Если в результате все пути окажутся разорваны, значит, множества переменных действительно независимы; если нет – нет.

Пример

Предположим, что может быть две причины, по которым трава может стать мокрой (GRASS WET): сработала дождевальная установка (SPRINKLER), либо прошёл дождь (RAIN). Также предположим, что дождь влияет на работу дождевальной машины (во время дождя установка не включается). Тогда ситуация может быть смоделирована проиллюстрированной Байесовской сетью. Каждая из трёх переменных может принимать лишь одно из двух возможных значений: T (правда — true) и F (ложь — false), с вероятностями, указанными в таблицах на иллюстрации.

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Совместная вероятность функции:

$$\mathrm{P}(G, S, R) = \mathrm{P}(G|S, R)\mathrm{P}(S|R)\mathrm{P}(R)$$

где имена трёх переменных означают G = Трава мокрая (Grass wet), S = Дождевальная установка (Sprinkler), и R = Дождь (Rain).

Модель может ответить на такие вопросы как «Какова вероятность того, что прошел дождь, если трава мокрая?» используя формулу условной вероятности и суммируя переменные:

$$\mathrm{P}(R = T \mid G = T) = \frac{\mathrm{P}(G = T, R = T)}{\mathrm{P}(G = T)} = \frac{\sum_{S \in \{T,F\}} \mathrm{P}(G = T, S, R = T)}{\sum_{S,R \in \{T,F\}} \mathrm{P}(G = T, S, R)}$$

$$= \frac{(0.99 \times 0.01 \times 0.2 = 0.00198_{TTT}) + (0.8 \times 0.99 \times 0.2 = 0.1584_{TFT})}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0_{TFF}} \approx 35.77\%.$$