Условная вероятность

$$P(B|A) = \frac{N(AB)}{N(A)} = \frac{N(AB)/N}{N(A)/N} = \frac{P(AB)}{P(A)}$$

N(AB) – число исходов, благоприятствующих совместному осуществлению A и B,
N(A) - число исходов, благоприятствующих осуществлению A

Условной вероятностью события B при условии A называется отношение вероятности совместного появления событий A и B к вероятности события A:

$$P(B|A) = \frac{P(AB)}{P(A)}, P(A) \neq 0$$

Условная вероятность

Другая запись условной вероятности:

$$P(B|A) = \boldsymbol{P_A(B)} = \frac{P(AB)}{P(A)}, P(A) \neq 0$$

$$P(A|H_1) = \boldsymbol{P_{H1}(A)}$$

Независимые события

События A и B называются **независимыми**, если

$$P(AB) = P(A) * P(B), \text{то есть } P(B|A) = P(B)$$

Формула умножения вероятностей (для конечного числа событий)

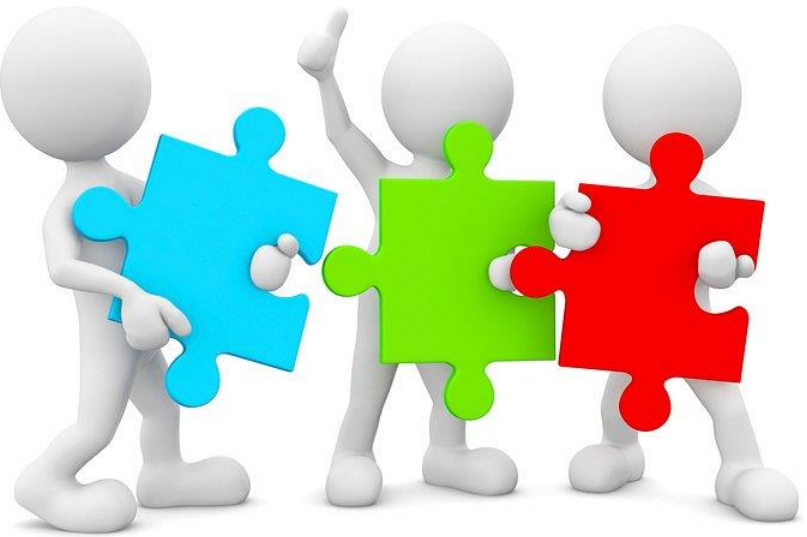
$$P(A_1 A_2 \dots A_n) = P(A_1) * P(A_2|A_1) * \dots * P(A_n|A_1\ A_2 \dots A_{n-1})$$

Полная вероятность

Предположим, что в результате испытания событие А может произойти одновременно с одним из попарно независимых событий H_1, H_2 … H_n, составляющих полную группу.

Тогда вероятность события А определяется **формулой полной вероятности:**

$$P(A) = \sum_{k=1}^{n} P(H_k) * P(A|H_k)$$



События H_1, H_2 … H_n называются гипотезами и удовлетворяют условию

$$\sum_{k=1}^{n} P(H_k) = 1$$

Полная вероятность

Пример: в кассе продается 70 билетов лотереи «Один из ста» и 30 билетов «Три из пятидесяти». Какова вероятность выиграть в лотерею, если купить только 1 билет?

A – выигрыш в лотерею

H_1 – купить билет лотереи «Один из ста»

H_2 – купить билет лотереи «Три из пятидесяти»

Полная вероятность

Пример: в кассе продается 70 билетов лотереи «Один из ста» и 30 билетов «Три из пятидесяти». Какова вероятность выиграть в лотерею, если купить только 1 билет?

A – выигрыш в лотерею

H_1 – купить билет лотереи «Один из ста»
H_2 – купить билет лотереи «Три из пятидесяти»

$P(H_1) = 70/100 = 0.7$	$P(A\|H_1) = 1/100 = 0.01$
$P(H_2) = 30/100 = 0.3$	$P(A\|H_2) = 3/50 = 0.06$

$$P(A) = P(H_1) * P(A|H_1) + P(H_2) * P(A|H_2) =$$
$$0.7 * 0.01 + 0.3 * 0.06 = 0.025$$

Полная вероятность

A – выигрыш в лотерею

H_1 – купить билет лотереи «Один из ста»
H_2 – купить билет лотереи «Три из пятидесяти»

} *гипотезы*

априорные (доопытные) вероятности

$P(H_1) = 70/100 = 0.7$	$P(A\|H_1) = 1/100 = 0.01$
$P(H_2) = 30/100 = 0.3$	$P(A\|H_2) = 3/50 = 0.06$

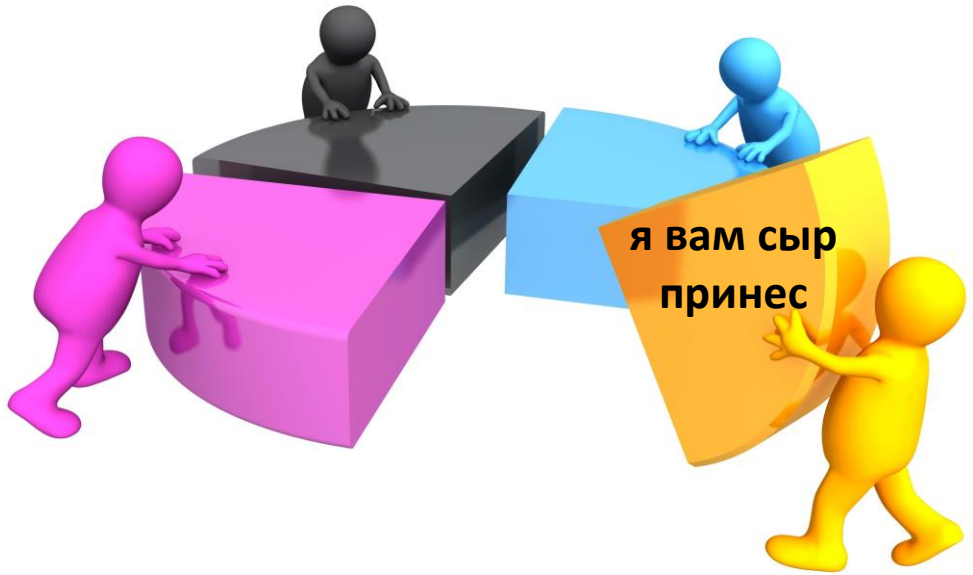
условные вероятности

$$P(A) = P(H_1) * P(A|H_1) + P(H_2) * P(A|H_2) =$$
$$0.7 * 0.01 + 0.3 * 0.06 = 0.025$$

Полная вероятность

Пример: Команда из четырех студентов сдает совместный проект. Зачет по проекту зависит от защиты, причем вызвать на защиту проекта могут любого студента из команды. Первый студент знает содержание проекта на 100 %, второй и третий – на 50%, а четвертый – на 2%. Какова вероятность, что проект будет сдан?

Полная вероятность



$$P(H_1) = P(H_2) = P(H_3) = P(H_4) = 0.25$$

$$P(A|H_1) = 1$$
$$P(A|H_2) = 0.5$$
$$P(A|H_3) = 0.5$$
$$P(A|H_4) = 0.02$$

$$\boldsymbol{P(A)} = P(H_1) * P(A|H_1) + P(H_2) * P(A|H_2) +$$
$$+ \ P(H_3) * P(A|H_3) + \ P(H_4) * P(A|H_4) =$$
$$0.25 * 1 + 0.25 * 0.5 + 0.25 * 0.5 + 0.25 * 0.02 = \mathbf{0.505}$$

Полная вероятность

Пример:

Вероятности того, что во время работы цифровой электронной машины произойдет сбой в арифметическом устройстве, в оперативной памяти, в остальных устройствах, относятся как *3:2:5.*

Вероятности обнаружения сбоя в арифметическом устройстве, в оперативной памяти и в остальных устройствах соответственно равны *0,8; 0,9; 0,9.*

Найти вероятность того, что возникший в машине сбой будет обнаружен.

Полная вероятность

A – возникновение сбоя в машине

H_1 – сбой в арифметическом устройстве

H_2 – сбой в оперативной памяти

H_3 – сбой в других устройствах

Априорные вероятности:

$P(H_1) = ?$

$P(H_2) = ?$

$P(H_3) = ?$

Условные вероятности:

$P(A|H_1) = ?$

$P(A|H_2) = ?$

$P(A|H_3) = ?$

Полная вероятность

A – возникновение сбоя в машине

H_1 – сбой в арифметическом устройстве

H_2 – сбой в оперативной памяти

H_3 – сбой в других устройствах

Априорные вероятности:

$P(H_1) = 0.3$

$P(H_2) = 0.2$

$P(H_3) = 0.5$

$\Sigma P(H_i) = 1$

Условные вероятности:

$P(A|H_1) = 0.8$

$P(A|H_2) = 0.9$

$P(A|H_3) = 0.9$

Полная вероятность

Полная вероятность:

$$P(A) = P(A|H_1) * P(H_1) + P(A|H_2) * P(H_2) + P(A|H_3) * P(H_3)$$

$$= 0.8 * 0.3 + 0.9 * 0.2 + 0.9 * 0.5 = 0.87$$

Априорные вероятности:

$P(H_1) = 0.3$

$P(H_2) = 0.2$

$P(H_3) = 0.5$

Условные вероятности:

$P(A|H_1) = 0.8$

$P(A|H_2) = 0.9$

$P(A|H_3) = 0.9$

Формула Байеса

Пример:

В цифровой электронной машине произошел сбой.

Какова вероятность, что причиной стал сбой в арифметическом устройстве, в оперативной памяти, в остальных устройствах?

Формула Байеса

Априорные вероятности:

$P(H_1) = 0.3$

$P(H_2) = 0.2$

$P(H_3) = 0.5$

Полная вероятность:

$P(A) = 0.87$

Условные вероятности:

$P(A|H_1) = 0.8$

$P(A|H_2) = 0.9$

$P(A|H_3) = 0.9$

Апостериорные (послеопытные) вероятности

$$P(H_1|A) = ?$$

$$P(H_2|A) = ?$$

$$P(H_3|A) = ?$$

Формула Байеса

Апостериорные (послеопытные) вероятности гипотез H_i после того как произошло событие A:

Формула Байеса (Бейеса)

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

$$P(\boldsymbol{H_i}|A) = \frac{P(A|H_i) * P(H_i)}{P(A)} = \frac{P(A|\boldsymbol{H_i}) * P(\boldsymbol{H_i})}{\sum_{k=1}^{n} P(A|H_k) * P(H_k)}$$

Формула Байеса

Априорные вероятности:

$P(H_1) = 0.3$

$P(H_2) = 0.2$

$P(H_3) = 0.5$

Условные вероятности:

$P(A|H_1) = 0.8$

$P(A|H_2) = 0.9$

$P(A|H_3) = 0.9$

Полная вероятность:

$P(A) = 0.87$

Апостериорные (послеопытные) вероятности

$$P(H_1|A) = \frac{P(A|H_1) * P(H_1)}{P(A)} = \frac{0.8 * 0.3}{0.87} = 0.28$$

Вероятность того, что причиной сбоя машины является сбой арифметического устройства

Формула Байеса

$$P(H_1|A) = \frac{P(A|H_1) * P(H_1)}{P(A)} = \frac{0.8 * 0.3}{0.87} = 0.28$$

Вероятность того, что причиной сбоя машины является сбой арифметического устройства

$$P(H_2|A) = \frac{P(A|H_2) * P(H_2)}{P(A)} = \frac{0.9 * 0.2}{0.87} = 0.21$$

Вероятность того, что причиной сбоя машины является сбой в оперативной памяти

$$P(H_3|A) = \frac{P(A|H_3) * P(H_3)}{P(A)} = \frac{0.9 * 0.5}{0.87} = 0.51$$

Вероятность того, что причиной сбоя машины является сбой в остальных устройствах

$$\sum P(H_i|A) = 1$$

Формула Байеса

Задача про котёнка: У котенка есть три любимых места для отдыха: на хозяйской подушке, в дедушкином тапке и в кресле хозяина дома, в которых его можно найти с равной вероятностью. Вероятность того, что котенка в течение 30 минут выгонят с первого места составляет 0.7, со второго – 0.8, с третьего – 0.5.

Котенок успел проспать всего 10 минут и его прогнали с любимого места. Какова вероятность, что он устроился спать на хозяйской подушке?

Формула Байеса

События:

A – котенка выгнали

H_1 – котенок спал на хозяйской подушке

H_2 – котенок спал в дедушкином тапке

H_3 – котенок спал в кресле

Априорные вероятности:

$P(H_1) = P(H_2) = P(H_3) = 1/3$

Условные вероятности:

$P(A|H_1) = 0.7$

$P(A|H_2) = 0.8$

$P(A|H_3) = 0.5$

Полная вероятность:

$P(A) = 0.7 * 1/3 + 0.8 * 1/3 + 0.5 * 1/3 = 2/3$

Апостериорная вероятность:

$P(H_1|A) = P(A|H_1) * P(H_1) / P(A) =$

$= 0.7 * (1/3) / (2/3) = 0.35$

Формула Байеса

Задача с повторением:

Имеется три партии деталей по 20 деталей в каждой. Число стандартных деталей в первой, второй и третьей партиях соответственно равны 20, 15, 10.

Из наудачу выбранной партии наудачу извлечена деталь, оказавшаяся стандартной. Деталь возвращают в партию и вторично из той же партии наудачу извлекают деталь, которая также оказывается стандартной.

Найти вероятность того, что детали были извлечены из третьей партии.

Формула Байеса

События

A – в каждом из двух испытаний (с возвращением) извлечена стандартная деталь

H_i – детали извлекались из i-й партии

Априорные вероятности

$P(H_1) = P(H_2) = P(H_3) = 1/3$

Условные вероятности

$P(A|H_1) = 1$

$P(A|H_2) = 15/20 * 15/20 = 9/16$

$P(A|H_3) = 10/20 * 10/20 = 1/4$

Апостериорная вероятность (формула Байеса)

$$P(H_3|A) = \frac{P(H_3) * P(A|H_3)}{P(H_1) * P(A|H_1)+P(H_2) * P(A|H_2)+P(H_3) * P(A|H_3)} =$$

$$= \frac{1/3*1/4}{1/3*1+1/3*9/16+1/3*1/4} = 4/29$$

Парадокс Байеса

Задача:

Существует заболевание с частотой распространения среди населения равна 0.001.

Существует метод, позволяющих обнаружить заболевание у больного с вероятностью 0.9.

Метод имеет вероятность ложноположительного результата — ошибочного выявления заболевания у здорового человека, которая равна 0.01.

Найти вероятность того, что человек здоров, если он был признан больным при обследовании.

Парадокс Байеса

События:

H – человек болен; $\overline{H}$ – человек здоров **гипотезы**

A – болезнь выявлена; $\overline{A}$ – человек признан здоровым **основное событие**

$P(H) = 0.001$ - вероятность заболевания

$P(\overline{H}) = 1 - P(H) = 0.999$ - вероятность того, что человек здоров

$P(A|H) = 0.9$ - вероятность выявить болезнь у больного

$P(A|\overline{H}) = 0.01$ - вероятность ошибочно выявить болезнь у здорового

$P(\overline{H}|A)$ - ?

Найти вероятность того, что человек здоров, если он был признан больным при обследовании.

Парадокс Байеса

$P(H) = 0.001$ - вероятность заболевания

$P(\bar{H}) = 1 - P(H) = 0.999$ - вероятность того, что человек здоров

$P(A|H) = 0.9$ - вероятность выявить болезнь у больного

$P(A|\bar{H}) = 0.01$ - вероятность ошибочно выявить болезнь у здорового

$$\mathbf{P(\bar{H}|A)} = \frac{P(\bar{H}) * P(A|\bar{H})}{P(\bar{H}) * P(A|\bar{H}) + P(H) * P(A|H)} =$$

$$= \frac{0.999 * 0.01}{0.999 * 0.01 + 0.001 * 0.9} = 0.917$$

Парадокс Байеса

H – человек болен; $\overline{H}$ – человек здоров

A – болезнь выявлена; $\overline{A}$ – человек признан здоровым

$P(\overline{H}|A)$ = 0.917

91,7 % людей, у которых на обследовании выявили заболевание, на самом деле здоровы.

Причина в том, что по условию задачи вероятность ложноположительного результата хоть и мала, но на порядок больше доли больных в обследуемой группе людей.

$$P(A|\overline{H}) = 0.01 \gg P(H) = 0.001$$

Парадокс Байеса

Повторное обследование того же человека будет давать независимый от первого результат. Имеет смысл провести повторное обследование людей, получивших результат «болен». Вероятность того, что человек здоров после получения повторного результата «болен», также можно вычислить по формуле Байеса.

B – болезнь выявлена дважды (при двух независимых обследованиях)

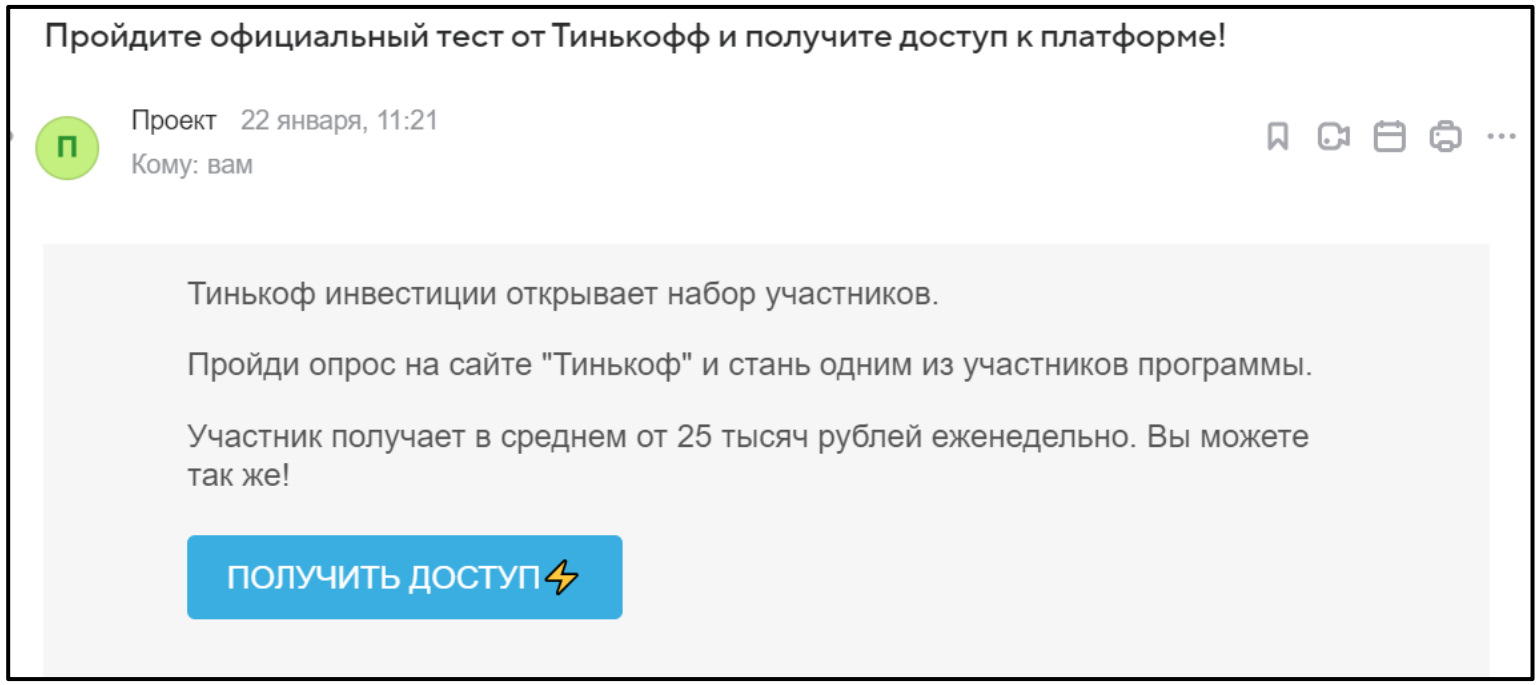
$$\mathbf{P(\bar{H}|B)} = \frac{P(\bar{H}) * P(B|\bar{H})}{P(\bar{H}) * P(B|\bar{H}) + P(H) * P(B|H)} =$$

$$= \frac{0.999 * 0.01 * 0.01}{0.999 * 0.01 * 0.01 + 0.001 * 0.9 * 0.9} = \mathbf{0.1098}$$

11 % людей, у которых на обследовании дважды выявили заболевание, на самом деле здоровы.

Байесовская фильтрация спама

Существует метод для фильтрации спама, основанный на применении **наивного байесовского классификатора**, в основе которого лежит применение теоремы Байеса.



Является ли письмо **спамом**?

Бинарный классификатор в машинном обучении

Object	Author	Content	Target/Class
Email 1	Проект	Тинькоф инвестиции открывает набор участников. Пройди опрос на сайте "Тинькоф" и стань одним из участников программы. Участник получает в среднем от 25 тысяч рублей еженедельно. Вы можете так же!	1 (Spam)
Email 2	Leader-ID	Этой весной Университет 2035 запускает 12-ю волну проектно-образовательного интенсива «От идеи к прототипу» — преакселератора для студенческих команд на начальном этапе знакомства с проектной деятельностью. Узнать подробнее о программе запуска проектно-образовательного интенсива «От идеи к прототипу» весны-24 вы сможете на вебинаре 14 февраля в 14:00 мск.	1 (Spam)
Email 3	Papa John's	Регистрация на сайте Papa John's Ваш пароль: *****	1 (Spam)
Email 4	Витрина подарков	Успейте заказать дебетовую карту для себя или близких до 31 марта и получите: кешбэк 25% за покупки на популярных маркетплейсах бесплатное обслуживание карты удобные платежи и переводы без комиссии кешбэк 10% за все остальные покупки	0 (Non - spam)

Наивный байесовский классификатор

- Первой известной программой, фильтрующей почту с использованием байесовского классификатора, была программа **iFile Джейсона Ренни**, выпущенная в 1996 году.
- В 2002 г. **Пол Грэм** смог значительно уменьшить число ложноположительных срабатываний до такой степени, что байесовский фильтр мог использоваться в качестве единственного фильтра спама.
- Модификации основного подхода были развиты во многих исследовательских работах и внедрены в программных продуктах. Многие современные почтовые клиенты осуществляют байесовское фильтрование спама.
- Фильтры для почтового сервера — такие, как **DSPAM, SpamAssassin, SpamBayes, SpamProbe, Bogofilter, CRM11**4 — используют методы байесовского фильтрования спама.

Наивный байесовский классификатор

Object	Author	Content	Target/Class
Email 2	Leader-ID	Этой весной Университет 2035 запускает 12-ю волну проектно-образовательного интенсива «От идеи к прототипу» — преакселератора для студенческих команд на начальном этапе знакомства с проектной деятельностью. Узнать подробнее о программе запуска проектно-образовательного интенсива «От идеи к прототипу» весны-24 вы сможете на вебинаре 14 февраля в 14:00 мск.	1 (Spam)

У каждого письма существует собственный «вес», **«спамовость» - вероятность оказаться спамом**. На основании этой вероятности принимается решение о том, считать ли письмо спамом.

«Спамовость» письма рассчитывается через «спамовость» или «вес» входящих в него слов. У каждого слова имеется собственная вероятность быть спамовым.

База данных классификатора

- Необходимо создать **базу данных** со словами и знаками (например, знак $, IP-адреса и домены и т.д.).
- В нее можно внести слова и знаки, содержащиеся в шаблоне спам-сообщения или допустимого сообщения.
- При обучении фильтра для каждого встреченного в письмах слова высчитывается и сохраняется его **«вес»** — оценка вероятности того, что письмо с этим словом — спам.

N	Word/symbol	Spam-probability
1	Привет	0.5
2	Кредит	0.91
3	Ипотека	0.89
4	Лекция	0.1

База данных классификатора

- Выполняется анализ **исходящей почты** конкретного пользователя и его уже **известного спама** — анализируются все слова и знаки, содержащиеся в этих сообщениях.

- Используется заранее созданная база, сформированная на основе сообщений пользователей (например, этого почтового агента)

База данных классификатора

- При обучении фильтра для каждого встреченного в письмах слова высчитывается и сохраняется его **«вес»** — оценка вероятности того, что письмо с этим словом — спам.
- В простейшем случае в качестве оценки используется частота: **«появлений в спаме / появлений всего».**
- В более сложных случаях возможна **предварительная обработка** текста: приведение слов в начальную форму, удаление служебных слов, *вычисление «веса» для целых фраз*, транслитерация и прочее.

«Спамовость» слова на примере «ипотеки»

F – слово «ипотека» встретилось в письме **основное событие**

S – письмо спам; **гипотеза**

$\bar{S}$ - письмо надежное **гипотеза**

Вероятность того, что письмо – спам, если в нем есть слово «ипотека»:

$$\mathbf{P(S|F)} = \frac{P(S) * P(F|S)}{P(\bar{S}) * P(F|\bar{S}) + P(S) * P(F|S)}$$

Большинство байесовских программ обнаружения спама работают как фильтры «без предубеждений», то есть полагают, что у любого сообщения равная вероятность являться и не являться спамом.

$\mathbf{P(\bar{S}) = P(S) = 0.5}$, поэтому $\mathbf{P(S|F)} = \frac{P(F|S)}{P(F|\bar{S}) + P(F|S)}$

«Спамовость» слова на примере «ипотеки»

Слово «ипотека» встречается:

- в 400 из 3000 спам-сообщений
- в 5 из 300 надежных сообщений

Значение его вероятности:

$$\mathbf{P(S|F)} = \frac{P(F|S)}{P(F|\bar{S}) + P(F|S)} = \frac{400 / 3000}{5 / 300 + 400 / 3000} = \mathbf{0.89}$$

«Спамовость» сообщения

F_i – присутствие i-го слова в сообщении

$P(S|F_i)$ – вероятность того, что письмо – спам,
если в нем содержится i-е слово

Вероятность «спамовости» всего сообщения,
содержащего слова F_1, … F_N:

$$\mathbf{P(S|F_1, \dots, F_N)} = \frac{P(S) * P(F_1, \dots, F_N|S)}{P(F_1, \dots, F_N)}$$

«Спамовость» сообщения

$$\mathbf{P(S|F_1, \dots, F_N)} = \frac{P(S) * P(F_1, \dots, F_N|S)}{P(F_1, \dots, F_N)} =$$

$$= [\text{предполагаем, что } F_i \text{ независимы}] =$$

$$= \frac{P(S) * P(F_1|S) * P(F_2|S) * \dots * P(F_N|S)}{P(F_1, \dots, F_N)} =$$

$$= \frac{P(S) * \prod_i^N P(F_i|S)}{P(F_1, \dots, F_N)} =$$

«Спамовость» сообщения

$$= \frac{P(S) * \prod_i^N P(F_i|S)}{P(F_1, \dots, FN)} =$$

= [по формуле полной вероятности] =

$$= \frac{P(S) * \prod_i^N P(F_i|S)}{\prod_i^N P(F_i|S) * P(S) + \prod_i^N P(F_i|\bar{S}) * P(\bar{S})} = \text{[преобразования]} =$$

$$\frac{\prod_i^N P(S|F_i)}{\prod_i^N P(S|F_i) + \prod_i^N (1 - P(S|F_i))}$$

«Спамовость» сообщения

Вероятность «спамовости» всего сообщения, содержащего слова F_1, … F_N:

$$P(S|F_1, \dots, F_N) = \frac{\prod_i^N P(S|F_i)}{\prod_i^N P(S|F_i) + \prod_i^N (1 - P(S|F_i))}$$

Байесовская фильтрация спама

В зависимости от полученной вероятности сообщения делается вывод о том, считать ли сообщение спамом.

- Обычно сообщение признается спамом, если вероятность «спамовости» **более 50 %**.
- В некоторых случаях положительное решение может приниматься, если вероятность **более 60-80 %.**

Байесовский классификатор

Object	Author	Content	Spam-prob-ty	Target/ Class
Email 1	Проект	Тинькоф инвестиции открывает набор участников. Пройди опрос на сайте "Тинькоф" и стань одним из участников программы. Участник получает в среднем от 25 тысяч рублей еженедельно. Вы можете так же!	0.9	1 (Spam)
Email 2	Leader-ID	Этой весной Университет 2035 запускает 12-ю волну проектно-образовательного интенсива «От идеи к прототипу» — преакселератора для студенческих команд на начальном этапе знакомства с проектной деятельностью. Узнать подробнее о программе запуска проектно-образовательного интенсива «От идеи к прототипу» весны-24 вы сможете на вебинаре 14 февраля в 14:00 мск.	0.55	1 (Spam)
Email 3	Papa John's	Регистрация на сайте Papa John's Ваш пароль: *****	0.6	1 (Spam)
Email 4	Витрина подарков	Успейте заказать дебетовую карту для себя или близких до 31 марта и получите: кешбэк 25% за покупки на популярных маркетплейсах, бесплатное обслуживание карты, удобные платежи и переводы без комиссии, кешбэк 10% за все остальные покупки	0.4	0 (Non - spam)