НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

INTRODUCTION TO NAIVE BAYES CLASSIFICATION

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 05.10.2018

Различия в подходах к теории вероятностей

Случайная величина — это величина, которая принимает в результате опыта одно из множества значений, причём появление того или иного значения этой величины до её измерения нельзя точно предсказать.

1. В частотном подходе (классический подход) предполагается, что случайность есть объективная неопределенность. Вероятность рассчитывается из серии экспериментов и является мерой случайности как эмпирической данности. Исторически частотный подход возник из практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл.
2. В байесовском подходе предполагается, что случайность характеризует наше незнания. Например, случайность при бросании кости связана с незнанием динамических характеристик игральной кости, сопротивления воздуха и так далее.

Многие задачи частотным методом решить невозможно (точнее, вероятность искомого события строго равна нулю). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ.

Понятие вероятности

Вероятность события — Вероятностью события А называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов. Например. Вероятность того, что на кубике выпадет четное число, равна следующему отношению **P=3/6=1/2**.

Формулы умножения вероятностей

Пусть события А и В независимые, причем вероятности этих событий известны. Найдем вероятность совмещения событий А и В.
Теорема: Вероятность совместного появления двух независимых событий равна произведению вероятностей этих событий:
$\mathbf{P(AB)=P(A)\cdot P(B)}$

Следствие Вероятность совместного появления нескольких событий, независимых в совокупности, равна произведению вероятностей этих событий:
$\mathbf{P(A_1A_2\ldots A_n)=P(A_1)\cdot P(A_2)\cdot \ldots P(A_n)}$

Понятие условной вероятности

Условной вероятностью события **A** при условии, что произошло событие **B**, называется число **P(A|B)=P(B, A)/ P(B)**,

P(B, A) – произведение вероятностей,
P(B) – полная вероятность события **B.**

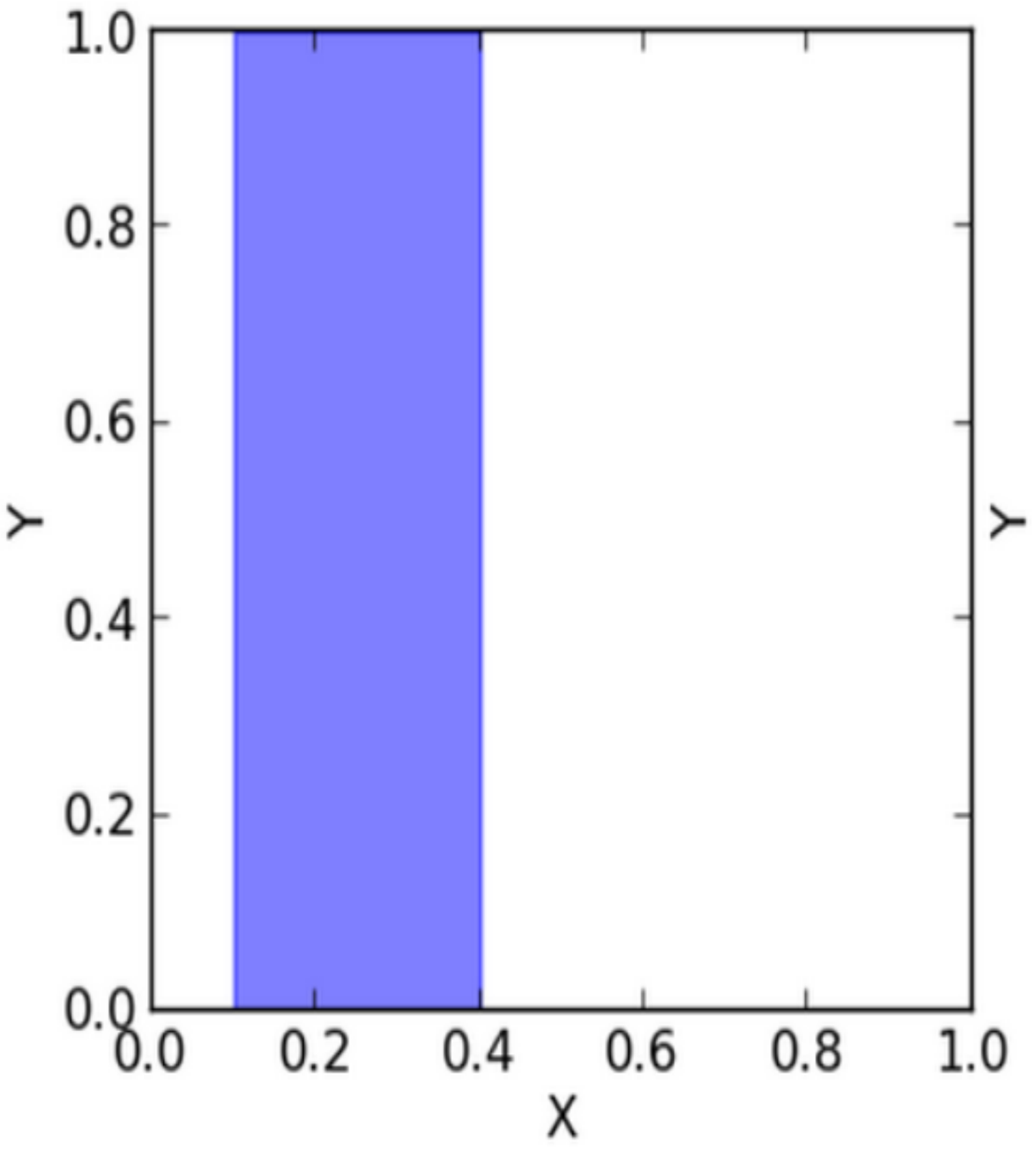
Например. В урне 3 белых и 3 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (**событие A**), если при первом испытании был извлечен черный шар (**событие B**).

Решение задачи:
Событие B – это вытаскивание первого шара (а именно черного). Вероятность события B=3/6=1/2 – вер. вытащить черный шар.
События A – это вытаскивание второго шара (а именно белого), так как в урне осталось 5 шаров, то вероятность этого события A=3/5
Таким образом, совместная вероятность событий A и B это произведение вероятностей этих событий P(B, A) =(3/6)*(3/5)=9/30
Полная вероятность события B=1/2

Итоговый результат: {3/6*3/5}/(1/2)=**3/5**

Геометрическая интерпретация вероятности

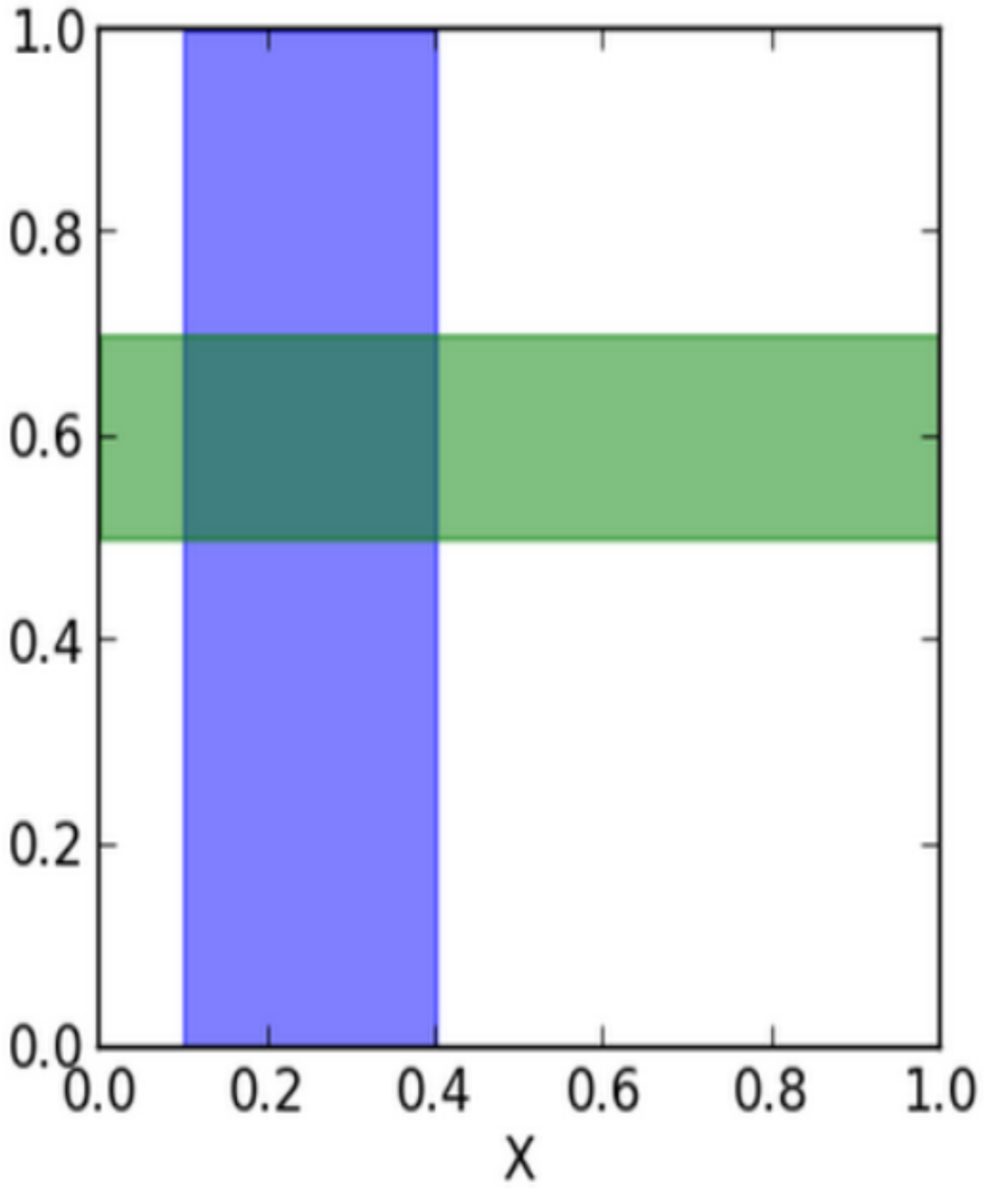
Рассмотрим следующий эксперимент: мы называем любое число из отрезка [0, 1] и смотрим за тем, что это число будет между, например, 0.1 и 0.4. Как нетрудно догадаться, вероятность этого события будет равна отношению длины отрезка [0.1, 0.4] к общей длине отрезка [0, 1] (другими словами, отношение «количества» возможных равновероятных значений к общему «количеству» значений), то есть (0.4 — 0.1) / (1 — 0) = 0.3, то есть вероятность попадания в отрезок [0.1, 0.4] равна 30%.



Геометрическая интерпретация вероятности

вероятность того, что y находится внутри отрезка [0.5, 0.7] равна отношению площади зеленой области к площади всего квадрата p(0.5 <= y <= 0.7) = 0.2, или для краткости **p(Y)** = 0.2.

А теперь допустим мы хотим знать какова вероятность того, что y находится в интервале [0.5, 0.7], если x уже находится в интервале [0.1, 0.4]. При условии независимости этих событий, мы можем записать эту вероятность как **произведение двух вероятностей, соответственно это будет площадь темной фигуры (пересечение синий и зеленой полосы)**

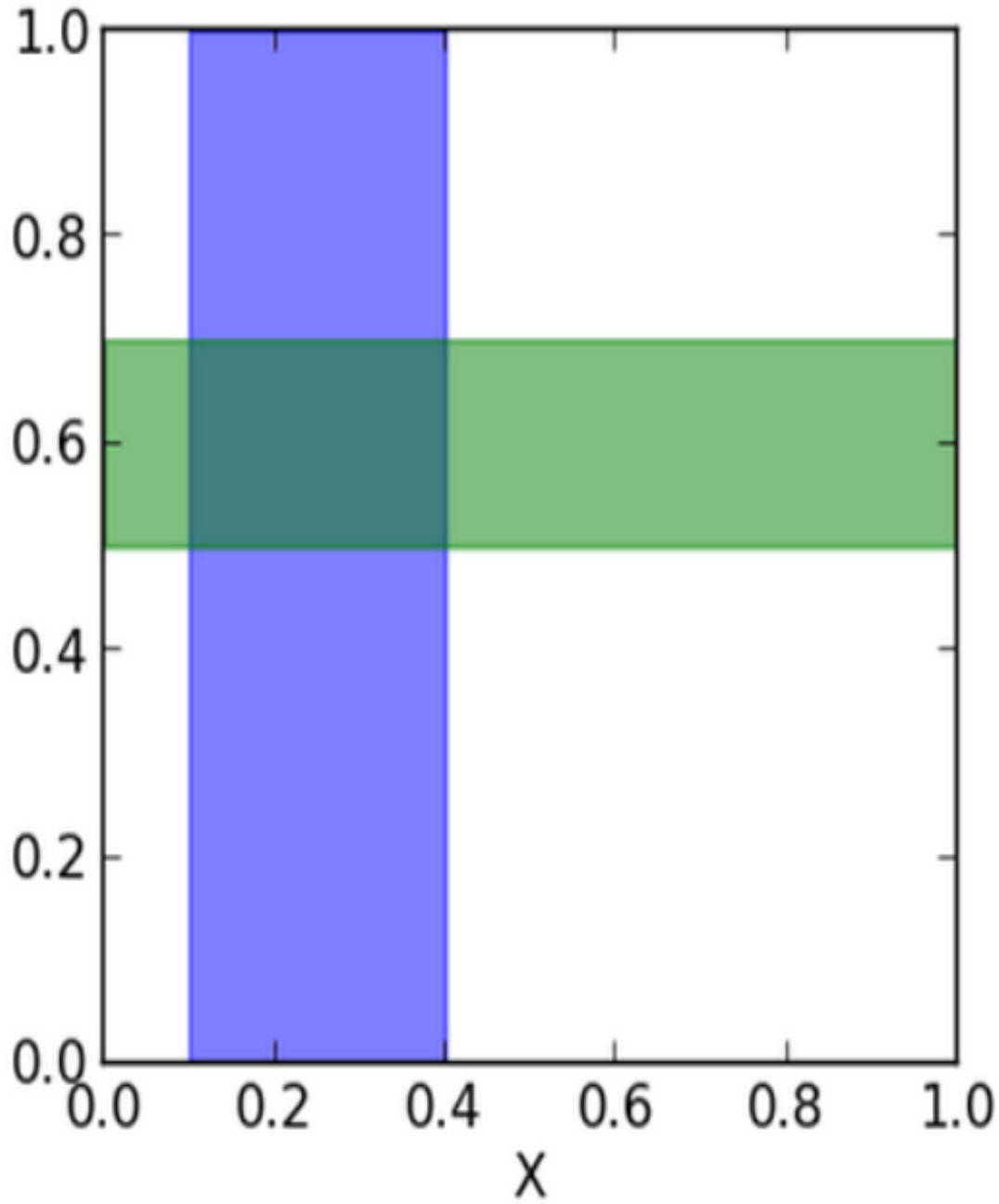


Геометрическая интерпретация вероятности

Таким полная вероятность того что наша точка попадает во внутрь закрашенной площади будет :
P= закрашенная площадь/общая площадь квадрата.

P= (0.2*0.3)/1

P(A|B)=P(B, A)/ P(B),



Байесовская вероятность

Байесовская вероятность — это интерпретация понятия вероятности, используемое в байесовской теории. Вероятность определяется как степень уверенности в истинности суждения**.**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

P(A) — априорная вероятность гипотезы *A (заранее известная вероятность)*;

P(A|B) — вероятность гипотезы *A* при наступлении события *B* **(апостериорная вероятность**);
P(B|A) — вероятность наступления события *B* при истинности гипотезы *A*;
P(B) — полная вероятность наступления события *B*.
P(A|B)) — вероятность наступления события *A* при истинности гипотезы *B*;

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Таким образом, формула Байеса может быть использована для разработки алгоритмов классификации.

Априорные и апостериорные суждения

1. Предположим, мы хотим узнать значение некоторой неизвестной величины.

2. У нас имеются некоторые знания, полученные до (a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания.

3. В процессе наблюдений эти знания подвергаются постепенному уточнению. После (a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении.

4.Будем считать, что мы пытаемся оценить неизвестное значение величины P(A|B) посредством наблюдений некоторых ее косвенных характеристик (гипотез).

5. В зависимости от уровня вероятности мы можем принять или отвергнуть нашу гипотезу (предсказание)

Если у нас много событий то мы предполагаем что они не зависят друг от друга. Например, мы считали что процесс вытаскивания шара из урны не зависит от цвета шара. В связи с таким допущением алгоритм называется «наивным».

Пример применения формулы Байеса в E-Health

Пример: случайному пациенту сделали тест на наличие СПИД, и получили положительный результат. Пусть точность теста 99.8% (т.е. он дает положительный результат у 0.2% здоровых людей). **Какова вероятность, что у этого пациента СПИД?**

Априорная вероятность **P(больной) =0.3** – доля больных в стране (пусть 0.3%)

P(тест +| больной) = 1 : тест дал положительный результат.

P(тест +| здоровый) = 1

P(больной) =0.2 – доля здоровых в стране (0.2%)

$$P = \frac{P(\text{тест} + |\ \text{больной}) \cdot P(\text{больной})}{P(\text{тест} + |\ \text{больной}) \cdot P(\text{больной}) + P(\text{тест} + |\ \text{здоровый}) \cdot P(\text{здоровый})}$$

$$P = \frac{1 \cdot 0.3}{1 \cdot 0.3 + 1 \cdot 0.2} = 0.6$$

Вероятностная постановка задачи классификации

Пусть имеется множество объектов X и конечное множество классов Y . Требуется построить алгоритм способный классифицировать произвольный объект X в рамках заданного множества Y.

Апостериорная вероятность принадлежности объекта X классу Y по формуле Байеса:

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

$P(A|B)$ - Апостериорная вероятность

$p(A,B)$ - Априорная вероятность

Задача классификации заключается в расчете (оценке) апостериорной информации на основании априорной информации. Такая оценка может быть реализована при помощи формулы Байеса. Однако существует проблема оценивания априорной величины p(A,B)

Задача восстановления априорного распределения

p(A,B)

Оценка функции ***p(A,B)*** может быть реализован при помощи трех методов.

1. Непараметрическое восстановление плотности основано на локальной аппроксимации плотности p(x) в окрестности классифицируемого объекта x ∈ X. Пример, Алгоритм Парзена-Розенблатта (метод парзеновского окна).
2. Параметрическое восстановление плотности основано на предположении, что плотность распределения известна с точностью до параметра, ***p(x,y)*** = ϕ(x; θ), где ϕ фиксированная функция. Пример. Нормальный дискриминантный анализ. LSA – в основе лежит метод SVD разложения.
3. Восстановление смеси плотностей. Если функцию плотности ***p(x,y)*** не удаётся смоделировать параметрическим распределением, можно попытаться описать её смесью нескольких распределений:

Собственно именно третий метод является основой тематического моделирования

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x; \theta_j), \quad \sum_{j=1}^{k} w_j = 1,$$

Как работает наивный байесовский алгоритм?

Пусть у нас есть набор данных, содержащий один признак «Погодные условия» (weather) и целевую переменную «Игра» (play), которая обозначает возможность проведения матча. На основе погодных условий мы должны определить (предсказать), состоится ли матч.

1. Пусть у нас есть набор наблюдений

	A	B
1	weather	play
2	sunny	no
3	overcast	yes
4	rainy	yes
5	sunny	yes
6	sunny	yes
7	overcast	yes
8	rainy	no
9	rainy	no
10	sunny	yes
11	rainy	yes
12	sunny	no
13	overcast	yes
14	overcast	yes
15	rainy	no

2. Преобразуем данные в таблицу частот

A	B	C
weather	no	yes
overcast	0	4
rainy	3	2
sunny	2	3
total	5	9

3. Преобразуем частоты в таблицу вероятности

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Как работает наивный байесовский алгоритм?

Задача: какова вероятность проведения матча в зависимости от погоды.

Решение: X – это **Да** или **Нет** (то есть у нас два класса).
C – типы погоды (overcast, sunny, rainy) - признаки

Вероятность проведения матча в солнечную погоду.

По формуле Байеса:

$$P(yes/sunny) = \frac{\mathrm{P}(\mathrm{sunny}/yes) * \mathrm{P}(\mathrm{yes})}{P(sunny)}$$

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

$$\mathrm{P}(\mathrm{sunny}/yes) = 3/9$$

$$\mathrm{P}(\mathrm{yes}) = 9/14$$

$$P(sunny) = 5/14$$

$$P(yes/sunn) = \frac{\left(\frac{3}{9}\right) * \left(\frac{9}{14}\right)}{P\left(\frac{5}{14}\right)} = 0.6$$

$$P(no/sunny) = \frac{\mathrm{P}(\mathrm{sunny}/no) * \mathrm{P}(\mathrm{no})}{P(sunny)}$$

$$\mathrm{P}(\mathrm{sunny}/no) = 2/5 \qquad P(sunny) = 5/14$$

$$\mathrm{P}(\mathrm{no}) = 5/14$$

$$P(no/sunn) = \frac{\left(\frac{2}{5}\right) * \left(\frac{5}{14}\right)}{P\left(\frac{5}{14}\right)} = 0.4$$

Пример оценки надежности компании

Пусть нам нужно оценить надежность компании. Мы предполагаем, что у нас есть три гипотезы о надежности ($Pr(\theta_{i:1,2,3})$). 1. Средняя надежность. 2. Высокая надежность.

3. Низкая надежность.

Номер гипотезы i	Средняя надежность (Pr1)	Высокая надежность (Pr2)	Низкая надежность (Pr3)
$Pr(\theta_i)$ (число компаний имеющих разные уровни надежности)	0.5 (50%)	0.3 (30%)	0.2 (20%)
Число компаний имеющие прибыл $Pr(y_1; \theta_i)$	0.4 (40%)	0.8 (80%)	0.3 (30%)
Число компаний, осуществляющие своевременный расчет с гос. $Pr(y_2; \theta_i)$	0.7 (70%)	0.9 (90%)	0(0%)

Вопрос, как будут меняться вероятности гипотез (Pr1, Pr2, P3) если мы наблюдаем какую либо величину? Расчет вероятности гипотез ведется при помощи формулы Байеса.

Пример оценки надежности компании

Пусть мы наблюдаем компанию у которой есть прибыль. Тогда гипотеза (апостериорное значение) того, что данная компания относится к типу средней надежности будет рассчитываться следующим образом.

$$Pr1 = \frac{0.4 * 0.5}{0.4 * 0.5 + 0.8 * 0.3 + 0.3 * 0.2} = 0.4 \text{ (было 0.5)}$$

Вероятность гипотезы о высокой надежности:

$$Pr2 = \frac{0.8*0.3}{0.4*0.5+0.8*0.3+0.3*0.2} = 0.48 \text{ (было 0.3)}$$

Вероятность гипотезы о низкой надежности:

$$Pr3 = \frac{0.3 * 0.2}{0.4 * 0.5 + 0.8 * 0.3 + 0.3 * 0.2} = 0.12 \text{(было 0.2)}$$

Таким образом мы получили апостериорные оценки, которые потом можно использовать как априорные.

Пример оценки надежности компании

Предположим, что фирма, которая имеет прибыль, еще и платит своевременно долги.

Номер гипотезы i	Средняя надежность	Высокая надежность	Низкая надежность
$Pr(\theta_i)$ (число компаний имеющих разные уровни надежности)	0.4	0.48	0.12
Число компаний имеющие прибыл $Pr(y_1; \theta_i)$	0.4	0.48	0.12
Число компаний, осуществляющие своевременный расчет с гос. $Pr(y_2; \theta_i)$	0.7	0.9	0

Тогда новые вероятности гипотез рассчитываются на основании предыдущих расчетов.

Средняя надежность $$Pr1 = \frac{0.4 * 0.7}{0.7 * 0.4 + 0.48 * 0.9 + 0 * 0.12} = 0.39 \text{ (было 0.4)}$$

Высокая надежность $$Pr2 = \frac{0.48 * 0.9}{0.7 * 0.4 + 0.48 * 0.9 + 0 * 0.12} = 0.607 \text{(было 0.48)}$$

Низкая надежность $$Pr3 = \frac{0.12 * 0}{0.7 * 0.4 + 0.48 * 0.9 + 0 * 0.12} = 0 \text{(было 0.12)}$$

Плюсы и минусы наивного байесовского алгоритма

Положительные стороны:

1. Классификация, в том числе многоклассовая, выполняется легко и быстро.
2. НБА лучше работает с категорийными признаками, чем с непрерывными.

Отрицательные стороны:

1. Если в тестовом наборе данных присутствует некоторое значение категорийного признака, которое не встречалось в обучающем наборе данных, тогда модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота».

2. Еще одним ограничением НБА является допущение о независимости признаков. В реальности наборы полностью независимых признаков встречаются крайне редко.

Области применения

1. Классификация в режиме реального времени. НБА очень быстро обучается, поэтому его можно использовать для обработки данных в режиме реального времени (котировки акций).
2. Многоклассовая классификация. НБА обеспечивает возможность многоклассовой классификации.
3. Классификация текстов, фильтрация спама, анализ тональности текста. При решении задач, связанных с классификацией текстов, НБА превосходит многие другие алгоритмы. Благодаря этому, данный алгоритм находит широкое применение в области фильтрации спама (идентификация спама в электронных письмах) и анализа тональности текста (анализ социальных медиа, идентификация позитивных и негативных мнений клиентов).

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru