Лекция 2. Байесовский подход к теории вероятностей. Примеры байесовских рассуждений

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Характеристики центра и разброса случайной величины

- Рассмотрим случайную величину X, имеющую плотность $p(x)$
- Математическое ожидание $\mathbb{E}X = \int xp(x)dx$ задает «центр тяжести» случайной величины, т.е. ее характерное среднее значение
- Средний квадрат отклонения значений случайной величины от ее мат. ожидания называется дисперсией

$$\mathbb{D}X = \mathbb{E}(X - \mathbb{E}X)^2 = \int (x - \mathbb{E}X)^2 p(x)dx$$

- Часто используются робастные (более устойчивые к выбросам) аналоги: медиана $\mathrm{med}\ X: \int_{-\infty}^{medX} p(x)dx = \int_{medX}^{+\infty} p(x)dx = 0.5$ и медиана абсолютных отклонений $\mathrm{med}\ |X - \mathrm{med}\ X|$

Условная вероятность

- Пусть X и Y — случайные величины с плотностями $p(x)$ и $p(y)$ соответственно
- В общем случае их совместная плотность $p(x,y) \neq p(x)p(y)$. Если это равенство выполняется, величины называют независимыми
- Условной плотностью называется величина
$$p(x|y) = \frac{p(x,y)}{p(y)}$$
- Смысл: как факт $Y = y$ влияет на распределение X. Заметим, что $\int p(x|y)dx \equiv 1$, но $\int p(x|y)dy$ не обязан равняться единице, т.к. относительно y это не плотность, а функция правдоподобия
- Очевидная система тождеств $p(x|y)p(y) = p(x,y) = p(y|x)p(x)$ позволяет легко переходить от $p(x|y)$ к $p(y|x)$
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Правило суммирования

- Все операции над вероятностями базируются на применении всего двух правил
- Правило суммирования: Пусть $A_1, \dots, A_k$ — взаимоисключающие события, одно из которых всегда происходит. Тогда

$$P(A_i \cup A_j) = P(A_i) + P(A_j) \qquad \sum_{i=1}^{k} P(A_i) = 1$$

- Очевидное следствие (формула полной вероятности): $\forall B$ верно $\sum_{i=1}^{k} P(A_i|B) = 1$, откуда

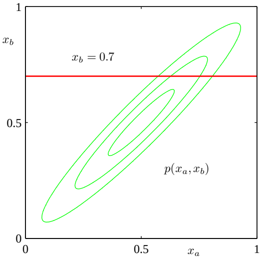
$$\sum_{i=1}^{k} \frac{P(B|A_i)P(A_i)}{P(B)} = 1 \qquad P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

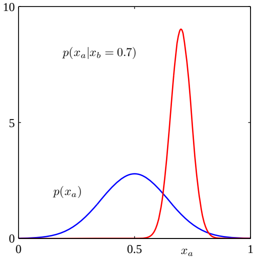
- В интегральной форме

$$p(b) = \int p(b,a)da = \int p(b|a)p(a)da$$

Условное и маргинальное распределения

- Плотность распределения интересующей нас компоненты x_a многомерной случайной величины (x_a, x_b) можно получить двумя способами в зависимости от имеющейся информации
- Если нам неизвестны значения остальных компонент, мы маргинализуем плотность по ним: $p(x_a) = \int p(x_a, x_b)dx_b$
- Если значения других компонент нам известны, то мы обуславливаем плотность по ним: $p(x_a|x_b) = \frac{p(x_a,x_b)}{p(x_b)}$





Правило произведения

- Правило произведения (product rule) гласит, что любую совместную плотность всегда можно разбить на множители

$$p(a,b) = p(a|b)p(b) \qquad P(A,B) = P(A|B)P(B)$$

- Аналогично для многомерных совместных распределений

$$p(a_1,\dots,a_n) = p(a_1|a_2,\dots,a_n)p(a_2|a_3,\dots,a_n)\dots p(a_{n-1}|a_n)p(a_n)$$

- Можно показать (Jaynes, 1995), что Sum- и Product-rule являются единственными возможными операциями, позволяющими рассматривать вероятности как промежуточную ступень между истиной и ложью

Ковариация и корреляция случайных величин

- Пусть X и Y — случайные величины с плотностями $p(x)$ и $p(y)$ соответственно
- Ковариацией X и Y называется следующая величина

$$Cov(X,Y) = \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y) = \int (x-\mathbb{E}X)(y-\mathbb{E}Y)p(x,y)dxdy$$

- Смысл ковариации: как отклонение одной случайной величины от своего мат. ожидания влияет на отклонение другой величины от своего мат. ожидания
- Легко показать (Упр.), что ковариация удовлетворяет всем аксиомам скалярного произведения
- Величина

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{D}X\mathbb{D}Y}}$$

называется коэффициентом корреляции двух случайных величин и определяет меру их линейной зависимости

Ковариационная матрица

- Пусть X является d-мерной случайной величиной с плотностью распределения $p(\boldsymbol{x})$
- Матрица
$$C = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T$$
назвается ковариационной матрицей. Ее элементы — это взаимные ковариации соответствующих компонент случайной величины, а на диагоналях стоят дисперсии $\mathbb{D}X_i$
- Аналогично, матрица корреляций содержит коэффициенты корреляций между соответствующими компонентами

Априорные и апостериорные суждения

- Предположим, мы хотим узнать значение некоторой неизвестной величины
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$

Формула Байеса

- Знаменитая формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений
- Обозначим априорные знания о величине θ за $p(\theta)$
- В процессе наблюдений мы получаем серию значений $\boldsymbol{x} = (x_1, \ldots, x_n)$. При разных θ наблюдение выборки $\boldsymbol{x}$ более или менее вероятно и определяется значением правдоподобия $p(\boldsymbol{x}|\theta)$
- За счет наблюдений наши представления о значении θ меняются согласно формуле Байеса

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- Заметим, что знаменатель не зависит от θ и нужен исключительно для нормировки апостериорной плотности

Условная независимость случайных величин

- Случайные величины x и y называются условно независимыми от z, если

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

- Другими словами вся информация о взаимозависимостях между x и y содержится в z
- Заметим, что из безусловной независимости не следует условная и наоборот
- Основное свойство условно независимых случайных величин

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

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

Пример

- Рассмотрим следующую гипотетическую ситуацию: римские легионы во главе с императором атакуют вторгшихся варваров
- События «гибель императора» и «уничтожение Рима» не являются независимыми
- Однако, если нам дополнительно известен исход битвы с варварами, эти два события становятся независимыми
- В самом деле, если легионы битву проиграли, то судьба Рима мало зависит от того, был ли император убит в сражении

Различия в подходах к теории вероятностей

- В современной теории вероятностей существуют два основных подхода к тому, что называть случайностью
- В частотном подходе предполагается, что случайность есть объективная неопределенность

  В жизни «объективные» неопределенности практически не встречаются. Чуть ли не единственным примером может служить радиоактивный распад (во всяком случае, по современным представлениям)
- В байесовском подходе предполагается, что случайность есть мера нашего незнания

  Практически любой случайный процесс можно так интерпретировать. Например, случайность при бросании кости связана с незнанием динамических характеристик кубика, сукна, руки кидающего, сопротивления воздуха и т.п.

Следствие частотного подхода

- При интерпретации случайности как «объективной» неопределенности единственным возможным средством анализа является проведение серии испытаний
- При этом вероятность события интерпретируется как предел частоты наступления этого события в n испытаниях при $n \to \infty$
- Исторически частотный подход возник из весьма важной практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл

Особенности частотного подхода

- Величины четко делятся на случайные и детерминированные
- Теоретические результаты работают на практике при больших выборках, т.е. при $n \gg 1$
- В качестве оценок неизвестных параметров выступают точечные, реже интервальные оценки
- Основным методом статистического оценивания является метод максимального правдоподобия (Фишер, 1930ые гг.)

Альтернативный подход

- Далеко не всегда при оценке вероятности события удается провести серию испытаний.
- Пример: оцените вероятность того, что человеческая цивилизация может быть уничтожена метеоритной атакой
- Очевидно, что частотным методом задачу решить невозможно (точнее вероятность этого события строго равна нулю, ведь подобного еще не встречалось). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ
- Идея байесовского подхода заключается в переходе от априорных знаний (или точнее незнаний) к апостериорным с учетом наблюдаемых явлений

Особенности байесовского подхода

- Все величины и параметры считаются случайными
  Точное значение параметров распределения нам неизвестно, значит они случайны с точки зрения нашего незнания
- Байесовские методы работают даже при объеме выборки 0! В этом случае апостериорное распределение равно априорному
- В качестве оценок неизвестных параметров выступают апостериорные распределения, т.е. решить задачу оценивания некоторой величины, значит найти ее апостериорное распределение
- Основным инструментом является формула Байеса, а также правила суммирования и произведения вероятностей

Недостатки байесовского подхода

- Начиная с 1930 гг. байесовские методы подвергались резкой критике и практически не использовались по следующим причинам
  - В байесовских методах предполагается, что априорное распределение известно до начала наблюдений и не предлагается конструктивных способов его выбора
  - Принятие решения при использовании байесовских методов в нетривиальных случаях требует колоссальных вычислительных затрат, связанных с численным интегрированием в многомерных пространствах
  - Фишером была показана оптимальность метода максимального правдоподобия, а следовательно — бессмысленность попыток придумать что-то лучшее
- В настоящее время (с начала 1990 гг.) наблюдается возрождение байесовских методов, которые оказались в состоянии решить многие серьезные проблемы статистики и машинного обучения

Точечные оценки при использовании метода Байеса

- Математическое ожидание по апостериорному распределению. Весьма трудоемкая процедура

$$\hat{\theta}_B = \int \theta p(\theta|\mathbf{x})d\theta$$

- Максимум апостериорной плотности. Удобен в вычислительном плане

$$\hat{\theta}_{MP} = \arg\max P(\theta|\mathbf{x}) = \arg\max P(\mathbf{x}|\theta)P(\theta) =$$

$$\arg\max \left(\log P(\mathbf{x}|\theta) + \log P(\theta)\right)$$

- Это коррекция оценки максимального правдоподобия

Попытки обобщения булевой логики

- Классическая булева логика плохо применима к жизненным ситуациям, которые далеко не всегда выразимы в терминах «истина» и «ложь»
- Неоднократно предпринимались попытки обобщить булеву логику, сохраняя при этом действие основных логических законов (Modus Ponens, Modus Tolens, правило де Моргана, закон двойного отрицания и пр.)
- Наиболее известные примеры:
    - Многозначная логика, расширившая множество логических переменных до $\{0, 1, \ldots, k-1\}$
    - Нечеткая логика, оперирующая континуумом значений между 0 и 1, характеризующими разную степень истинности

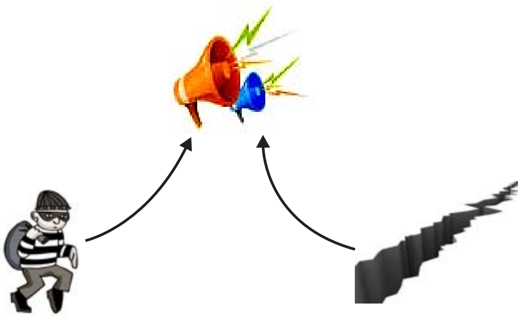
Недостатки нечеткой логики

- Несмотря на кажущуюся привлекательность нечеткая логика обладает рядом существенных недостатков
- Отсутствует строгое математическое обоснование для ряда методов, использующихся в нечетких рассуждениях
- Существует множество эвристических правил, определяющих как именно нужно строить нечеткий вывод. Все они приводят к различным результатам
- Непонятна связь нечеткой логики с теорией вероятности

Логическая интерпретация байесовского подхода

- Байесовский вывод можно рассматривать как обобщение классической булевой логики. Только вместо понятий «истина» и «ложь» вводится «истина с вероятностью p».
- Обобщение классического правила Modus Ponens
$$\frac{A, A \Rightarrow B}{A \& B} \qquad \frac{p(A), p(B|A)}{p(A \& B)}$$
- Теперь рассмотрим такую ситуацию
$$\frac{A \Rightarrow B, B}{A = ?} \qquad \frac{p(B|A), p(A)}{p(A|B)}$$
Формула Байеса позволяет рассчитать изменение степени истинности A с учетом информации о B
- Это новый подход к синтезу экспертных систем
- В отличие от нечеткой логики, он теоретически обоснован и математически корректен

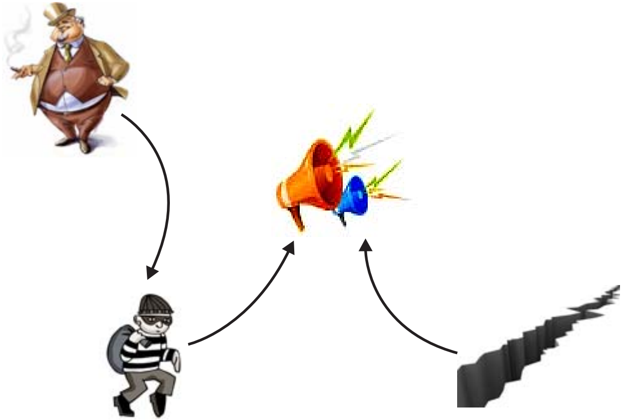
Жизненная ситуация



Вероятностная интерпретация

- Технические характеристики сигнализации $p(t|v,z) = p(t|v,\neg z) = 1$, $p(t|\neg v, z) = 0.1$, $p(t|\neg v,\neg z) = 0$
- Статистическая информация, набранная Джоном $p(v) = 2\cdot 10^{-4}$, $p(z) = 0.01$

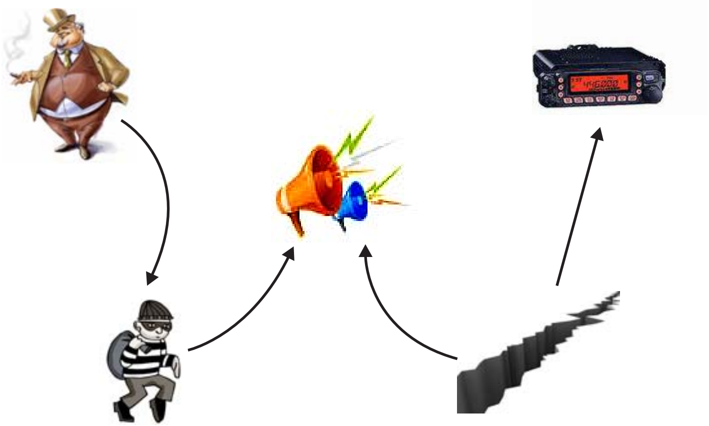
Жизненная ситуация



Вероятностная интерпретация

- Технические характеристики сигнализации $p(t|v,z) = p(t|v,\neg z) = 1$, $p(t|\neg v, z) = 0.1$, $p(t|\neg v, \neg z) = 0$
- Статистическая информация, набранная Джоном $p(v) = 2 \cdot 10^{-4}$, $p(z) = 0.01$
- Сообщение друга $p(d) = 1$, $p(v|d) = 2 \cdot 10^{-3}$, $p(v|\neg d) = 2 \cdot 10^{-4}$
- Мы предположим, что Джон полностью доверяет другу. Но мы легко могли бы учесть и тот факт, что друг Джона – большой шутник и мог его разыграть, положив $p(d) < 1$

Жизненная ситуация



Вероятностная интерпретация

- Технические характеристики сигнализации $p(t|v,z) = p(t|v,\neg z) = 1$, $p(t|\neg v, z) = 0.1$, $p(t|\neg v, \neg z) = 0$
- Статистическая информация, набранная Джоном $p(v) = 2 \cdot 10^{-4}$, $p(z) = 0.01$
- Сообщение друга $p(d) = 1$, $p(v|d) = 2 \cdot 10^{-3}$
- Сводка новостей по радио $p(r) = 1$, $p(r|z) = 0.5$, $p(r|\neg z) = 0$

Расчет вероятностей I

Срабатывание сигнализации: событие t

$$p(v|t) = \frac{1}{Z}p(t|v)p(v)$$

$$p(\neg v|t) = \frac{1}{Z}p(t|\neg v)p(\neg v)$$

$$Z = p(t|v)p(v) + p(t|\neg v)p(\neg v)$$

$$p(t|v) = p(t|v, \neg z)p(\neg z) + p(t|v, z)p(z) = p(\neg z) + p(z) = 1$$

$$p(t|\neg v) = p(t|\neg v, \neg z)p(\neg z) + p(t|\neg v, z)p(z) = p(t|\neg v, z)p(z) = 10^{-3}$$

$Z = 1.2 \cdot 10^{-3}$ $\qquad p(v|t) = \frac{1}{6} \approx 16.7\%$ $\qquad p(\neg v|t) = \frac{5}{6} \approx 83.3\%$

Расчет вероятностей II

Сообщение друга: событие d

$$p(v|t,d) = \{Cond.ind.\} = \frac{1}{Z}\frac{p(v|t)p(v|d)}{p(v)} = \frac{1}{Z}\frac{10}{6}$$

$$p(\neg v|t,d) = \{Cond.ind.\} = \frac{1}{Z}\frac{p(\neg v|t)p(\neg v|d)}{p(\neg v)} \approx \frac{1}{Z}\frac{5}{6}$$

$$Z = \frac{p(v|t)p(v|d)}{p(v)} + \frac{p(\neg v|t)p(\neg v|d)}{p(\neg v)}$$

$Z = \frac{15}{6}$ $\quad p(v|t,d) = \frac{10}{15} \approx 66.7\%$ $\quad p(\neg v|t,d) = \frac{5}{15} \approx 33.3\%$

Расчет вероятностей III

Радиосводка: событие r. Но т.к. $p(r|\neg z) = 0$, то $p(z|r) = 1$, т.е. имеет место землетрясение (событие z)

$$p(v|t,d,r) = \frac{1}{Z}p(t|v,r,d)p(v,r,d) = \frac{1}{Z}p(v,r,d) = \{Indep.assump.\} = \frac{1}{Z}p(v,d)p(r) = \frac{1}{Z}p(v|d)p(d)p(r) = \frac{1}{Z}2\cdot 10^{-3}\times 1\times 1$$

$$p(\neg v|t,d,r) = \left\{p(t|\neg v,d,r) = p(t|\neg v,d,z)p(z|r) + p(t|\neg v,d,\neg z)p(\neg z|r)\right\} = \frac{1}{Z}p(t|\neg v,r,d)p(\neg v,r,d) = \frac{1}{Z}0.1\times p(\neg v,r,d) = \{Indep.assump.\} = \frac{1}{Z}0.1\times p(\neg v,d)p(r) = \frac{1}{Z}0.1\times p(\neg v|d)p(d)p(r) = \frac{1}{Z}0.1\times(1-2\cdot 10^{-3})\times 1\times 1$$

$$Z = p(t|v,r,d)p(v,r,d) + p(t|\neg v,r,d)p(\neg v,r,d) = 0.1018$$

$$p(v|t,d,z) = \frac{20}{1018} \approx 1.9\%$$

$$p(\neg v|t,d,z) = \frac{998}{1018} \approx 98.1\%$$

Ошибка Джона

- Успокоенный Джон возвращается на работу, а вечером, придя домой, обнаруживает, что квартира «обчищена».
- Джон отлично владел байесовским аппаратом теории вероятностей, но значительно хуже разбирался в человеческой психологии
- Предположение о независимости кражи и землетрясения оказалось неверным

$$p(v, z) \neq p(v)p(z)$$

- Действительно, когда происходит землетрясение, воры проявляют значительно большую активность, достойную лучшего применения

$$p(v|z) > p(v|\neg z)$$