

Введение в машинное обучение и байесовский вывод



Машинное обучение
Основы байесовского вывода

01

Введение: история и структура AI и ML



Random facts:

- 4 сентября 476 г. предводитель варваров-наёмников Одоакр сверг римского императора Августула; так закончилась Западная Римская империя и началось королевство Одоакра, которое не просуществовало и двадцати лет
- 4 сентября 1607 г. произошло «бегство графов»: 99 ирландских аристократов, которые были потомками ирландских королей и вождей кланов, сбежали из Ирландии на кораблях в Испанию; из-за шторма они высадились во Франции, а затем направились в Рим, где получили убежище, но в Ирландию уже не вернулись
- 4 сентября 1837 г. в Нью-Йоркском университете Сэмюэл Морзе впервые продемонстрировал телеграф
- 4 сентября 1884 г. в ходе контрреформ Александра III в России была отменена автономия университетов
- 4 сентября 1975 г. в телеэфир вышел первый выпуск игры «Что? Где? Когда?»; ведущего не было, две семьи играли друг против друга, и раунды снимались у них дома
- 4 сентября 1998 г. была основана компания Google; она была расположена в Менло-Парке, в гараже Сьюзен Воджиски, подруги Ларри Пейджа и Сергея Брина



Первые мысли об искусственном интеллекте

- Гефест создавал себе роботов-андроидов, например гигантского человекоподобного робота Талоса.
- Пигмалион оживлял Галатею.
- Иегова и Аллах — куски глины.
- Особо мудрые раввины могли создавать големов.
- Альберт Великий изготовил искусственную говорящую голову (чем очень расстроил Фому Аквинского).

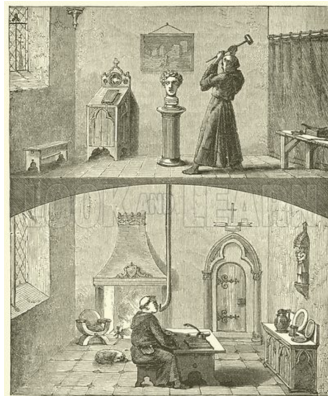


FIG. 432.—The Talking Head of Albertus Magnus.

Механические автоматы

- Роботы и автоматы вообще были весьма изобретательны: аль-Джазари (XII век)



- Замок Эден Роберта II д'Артуа, механический рыцарь Леонардо да Винчи...

Механические автоматы

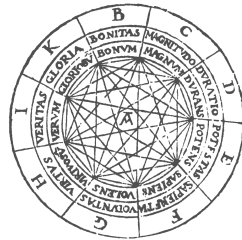


- Автоматоны Жаке-Дро (XVIII век):



Механические автоматы

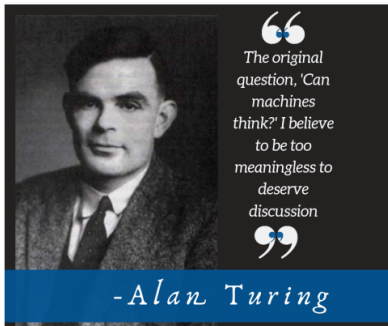
- Но это чистая механика,
а в математике/логике AI долго был
довольно рудиментарным
- Логическая машина Раймунда Луллия
(рубеж XIII-XIV вв.):



Начиная с доктора Франкенштейна, дальше AI в литературе появляется постоянно...

Тест Тьюринга

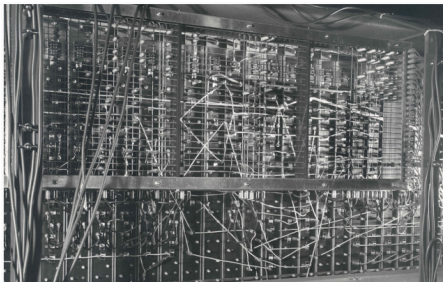
- AI как наука начался с **теста Тьюринга** (1950).
- Компьютер должен успешно выдать себя за человека в (письменном) диалоге между судьёй, человеком и компьютером. Правда, исходная формулировка была несколько тоньше и интереснее...



Тест Тьюринга



- Здесь уже очевидно, сколько всего надо, чтобы сделать AI:
 - обработка естественного языка;
 - представление знаний;
 - выводы из полученных знаний;
 - обучение на опыте (собственно machine learning).



Дартмутский семинар

- AI как наука появился в 1956 на семинаре в Дартмуте.
- Его организовали Джон Маккарти (John McCarthy), Марвин Мински (Marvin Minsky), Клод Шеннон (Claude Shannon) и Натаниэль Рочестер (Nathaniel Rochester).
- Это была, наверное, самая амбициозная грантозаявка в истории информатики.

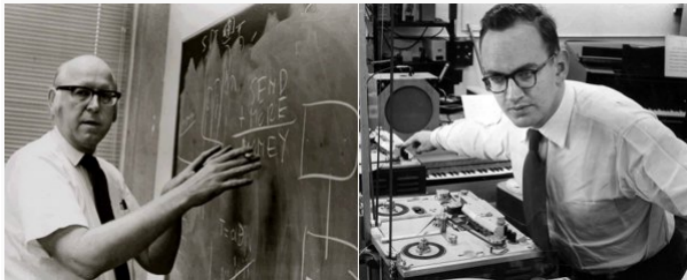


Дартмутский семинар

Мы предлагаем исследование искусственного интеллекта сроком в 2 месяца с участием 10 человек летом 1956 года в Дартмутском колледже, Гановер, Нью-Гемпшир. Исследование основано на предположении, что всякий аспект обучения или любое другое свойство интеллекта может в принципе быть столь точно описано, что машина сможет его симулировать. Мы попытаемся понять, как обучить машины использовать естественные языки, формировать абстракции и концепции, решать задачи, сейчас подвластные только людям, и улучшать самих себя. Мы считаем, что существенное продвижение в одной или более из этих проблем вполне возможно, если специально подобранная группа учёных будет работать над этим в течение лета.

1956-1960: большие надежды

- Оптимистическое время. Казалось, что ещё немного, ещё чуть-чуть...
- Allen Newell, Herbert Simon: **Logic Theorist**.
 - Программа для логического вывода.
 - Смогла передоказать большую часть **Principia Mathematica**, кое-где даже изящнее, чем сами Рассел с Уайтхедом.

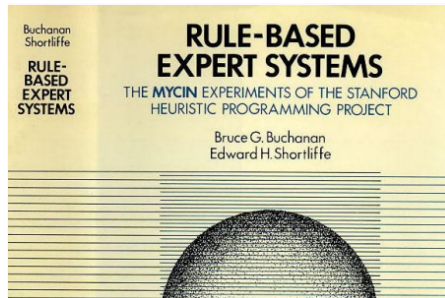


1956-1960: большие надежды

- Оптимистическое время. Казалось, что ещё немного, ещё чуть-чуть...
- General Problem Solver — программа, которая пыталась думать как человек;
- Много программ, которые умели делать некоторые ограниченные вещи (microworlds):
 - Analogy (IQ-тесты на «выберите лишнее»);
 - Student (алгебраические словесные задачи);
 - Blocks World (переставляла 3D-блоки).

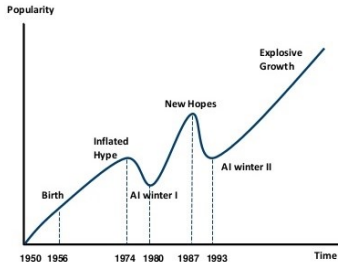
1970-e: knowledge-based systems

- Суть: накопить достаточно большой набор правил и знаний о предметной области, затем делать выводы.
- Первый успех: MYCIN – диагностика инфекций крови:
 - около 450 правил;
 - результаты как у опытного врача и существенно лучше, чем у начинающих врачей.



1980-е: коммерческие применения; индустрия AI

- Первый AI-отдел был в компании DEC (Digital Equipment Corporation); утверждают, что к 1986 году он сэкономил DEC около \$10 млн. в год.
- Бум закончился к концу 80-х, когда многие компании не смогли оправдать завышенных ожиданий.



Timeline of AI Development	
•	1950s-1960s: First AI boom - the age of reasoning, prototype AI developed
•	1970s: AI winter I
•	1980s-1990s: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
•	1990s: AI winter II
•	1997: Deep Blue beats Gary Kasparov
•	2006: University of Toronto develops Deep Learning
•	2011: IBM's Watson won Jeopardy
•	2016: Go software based on Deep Learning beats world's champions

1990-2010: data mining, machine learning



- В последние десятилетия основной акцент сместился на машинное обучение и поиск закономерностей в данных.
- Особенно — с развитием интернета.
- Сейчас про AI в смысле трёх законов робототехники уже не очень вспоминают.
- // Но роботика — процветает и пользуется machine learning на каждом шагу.

Определение



Что значит — обучающаяся машина? Как определить «обучаемость»?

Определение



Что значит — обучающаяся машина? Как определить «обучаемость»?



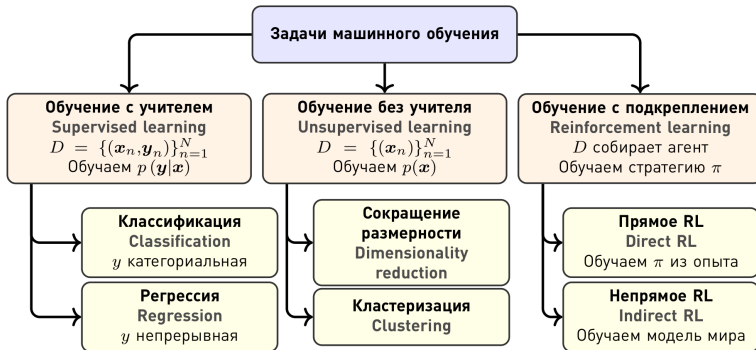
Определение. Компьютерная программа обучается по мере накопления опыта относительно некоторого класса задач T и целевой функции P , если качество решения этих задач (относительно P) улучшается с получением нового опыта.

- Определение очень (слишком?) общее.
- Какие конкретные примеры можно привести?

Чем мы будем заниматься

- Мы будем рассматривать разные алгоритмы, которые решают ту ли иную задачу, причём решают тем лучше, чем больше начальных (тестовых) данных ему дадут.
- Сегодня мы поговорим об общей теории байесовского вывода, в которую обычно можно погрузить любой алгоритм машинного обучения.
- Но сначала — краткий обзор основных задач машинного обучения в целом.

Основные задачи и понятия машинного обучения



Основные задачи и понятия машинного обучения

- **Обучение с учителем** (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами:
 - **обучающая выборка** (training set) – набор примеров, каждый из которых состоит из **признаков** (features, attributes);
 - у примеров есть правильные ответы – переменная (response), которую мы предсказываем; она может быть категориальная (categorical), непрерывная или ординальная (ordinal);

Основные задачи и понятия машинного обучения

- **Обучение с учителем** (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами:
- модель **обучается** на этой выборке (training phase, learning phase), затем может быть применена к новым примерам (test set);
 - главное – обучить модель, которая не только точки из обучающей выборки объясняет, но и на новые примеры хорошо **обобщается** (generalizes);
 - иначе – оверфиттинг (overfitting);

Основные задачи и понятия машинного обучения

- **Обучение с учителем** (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами:
- обычно нам дают просто обучающую выборку – как тогда проверить, обобщаются ли модели?
 - кросс-валидация – разбиваем выборку на тренировочный и валидационный набор (validation set);
 - перед тем как подавать что-то на вход, обычно делают предобработку, стараясь выделить из входных данных самые содержательные аспекты (feature extraction).

Основные задачи и понятия машинного обучения

- **Обучение с учителем** (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами:
 - **классификация:** есть некоторый дискретный набор категорий (классов), и надо новые примеры определить в какой-нибудь класс;
 - классификация текстов по темам, спам-фильтр;
 - распознавание лиц/объектов/текста;

Основные задачи и понятия машинного обучения

- **Обучение с учителем** (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами:
- **регрессия:** есть некоторая неизвестная функция, и надо предсказать её значения на новых примерах:
 - инженерные приложения (предсказать температуру, положение робота, whatever);
 - финансы – предсказать цену акций;
 - то же плюс изменения во времени – например, распознавание речи.

Основные задачи и понятия машинного обучения

- **Обучение без учителя** (unsupervised learning) – обучение, в котором нет правильных ответов, только данные:
- **кластеризация** (clustering): надо разбить данные на заранее неизвестные классы по некоторой мере похожести:
 - выделить семейства генов из последовательностей нуклеотидов;
 - кластеризовать пользователей и персонализировать под них приложение;
 - кластеризовать масс-спектрометрическое изображение на части с разным составом;

Основные задачи и понятия машинного обучения

- **Обучение без учителя** (unsupervised learning) – обучение, в котором нет правильных ответов, только данные:
- **снижение размерности** (dimensionality reduction): данные имеют огромную размерность (очень много признаков), нужно уменьшить её, выделить самые информативные признаки, чтобы все вышеописанные алгоритмы смогли работать;
 - **дополнение матриц** (matrix completion): есть разреженная матрица, надо предсказать, что на недостающих позициях.
 - Часто даны правильные ответы для небольшой части данных – semi-supervised learning.

Основные задачи и понятия машинного обучения

- **Обучение с подкреплением** (reinforcement learning) – обучение, в котором агент учится из собственных проб и ошибок:
- **многорукие бандиты**: есть некоторый набор действий, каждое из которых ведёт к случайным результатам; нужно получить как можно больший доход;
 - **exploration vs. exploitation**: как и когда от исследования нового переходить к использованию того, что уже изучил;
 - **credit assignment**: конфетку дают в самом конце (выиграл партию), и надо как-то распределить эту конфетку по всем ходам, которые привели к победе.

Основные задачи и понятия машинного обучения

- **активное обучение** (active learning) – как выбрать следующий (относительно дорогой) тест;
- **обучение ранжированию** (learning to rank) – ординальная регрессия, как породить упорядоченный список (интернет-поиск);
- **бустинг** (boosting) – как скомбинировать несколько слабых классификаторов так, чтобы получился хороший;
- **выбор модели** (model selection) – где провести черту между моделями с многими параметрами и с немногими.

Вероятность в машинном обучении

- Во всех методах и подходах очень пригодится метод, который мог бы не просто выдавать ответ, а ещё оценивать, насколько модель уверена в этом ответе, насколько модель хорошо описывает данные, как изменятся эти величины при дальнейших экспериментах и т.д.
- Поэтому центральную роль в машинном обучении играет теория вероятностей – и мы тоже будем её активно применять.

Источники

- Christopher M. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2007.
- Kevin Murphy, **Machine Learning: A Probabilistic Perspective**, MIT Press, 2013.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, 2nd ed., Springer, 2009.

Основные определения

- Нам не понадобятся математические определения сигма-алгебры, вероятностной меры, борелевских множеств и т.п.
- Достаточно понимать, что бывают дискретные случайные величины (неотрицательные вероятности исходов в сумме дают единицу) и непрерывные случайные величины (интеграл неотрицательной функции плотности равен единице).

Основные определения

- **Совместная вероятность** – вероятность одновременного наступления двух событий, $p(x, y)$; маргинализация:

$$p(x) = \sum_y p(x, y).$$

- **Условная вероятность** – вероятность наступления одного события, если известно, что произошло другое, $p(x | y)$:

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

- **Теорема Байеса** – из предыдущей формулы:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}.$$

- **Независимость**: x и y независимы, если

$$p(x, y) = p(x)p(y).$$

Прямые и обратные задачи

- Прямая задача: в урне лежат 10 шаров, из них 3 чёрных. Какова вероятность выбрать чёрный шар?
- Или: в урне лежат 10 шаров с номерами от 1 до 10. Какова вероятность того, что номера трёх последовательно выбранных шаров дадут в сумме 12?
- Обратная задача: перед нами две урны, в каждой по 10 шаров, но в одной 3 чёрных, а в другой — 6. Кто-то взял из какой-то урны шар, и он оказался чёрным. Насколько вероятно, что он брал шар из первой урны?
- Заметьте, что в обратной задаче вероятности сразу стали байесовскими (хоть здесь и можно переформулировать через частоты).

Прямые и обратные задачи

- Иначе говоря, прямые задачи теории вероятностей описывают некий вероятностный процесс или модель и просят подсчитать ту или иную вероятность (т.е. фактически по модели предсказать поведение).
- Обратные задачи содержат **скрытые переменные** (в примере — номер урны, из которой брали шар). Они часто просят по известному поведению построить вероятностную модель.
- Задачи машинного обучения обычно являются задачами второй категории.

Вероятность как частота

- Обычно в классической теории вероятностей, происходящей из физики, вероятность понимается как предел отношения количества определённого результата эксперимента к общему количеству экспериментов.
- Стандартный пример: бросание монетки.

Вероятность как степень доверия

- Мы можем рассуждать о том, «насколько вероятно» то, что
 - сборная России победит на чемпионате мира по футболу в 2026 году;
 - «Одиссею» написала женщина;
 - Александр Фёдорович Керенский бежал за границу в женском платье;
 - ...
- Но о «стремящемся к бесконечности количестве экспериментов» говорить бессмысленно — эксперимент здесь ровно один.

Вероятность как степень доверия

- Здесь вероятности уже выступают как **степени доверия** (degrees of belief). Это байесовский подход к вероятностям (Томас Байес так понимал).
- К счастью, и те, и другие вероятности подчиняются одним и тем же законам; есть результаты о том, что вполне естественные аксиомы вероятностной логики тут же приводят к весьма узкому классу функций (Сох, 19).

Определения

- Запишем теорему Байеса:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- Здесь $p(\theta)$ — **априорная вероятность** (prior probability), $p(D|\theta)$ — **правдоподобие** (likelihood), $p(\theta|D)$ — **апостериорная вероятность** (posterior probability), $p(D) = \int p(D | \theta)p(\theta)d\theta$ — **вероятность данных** (evidence).



ML vs. MAP

- В статистике обычно ищут **гипотезу максимального правдоподобия** (maximum likelihood):

$$\theta_{ML} = \arg \max_{\theta} p(D | \theta).$$

- В байесовском подходе ищут **апостериорное распределение** (posterior)

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

и, возможно, **максимальную апостериорную гипотезу** (maximum a posteriori):

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | D) = \arg \max_{\theta} p(D | \theta)p(\theta).$$



ЦЕНТРАЛЬНЫЙ
УНИВЕРСИТЕТ

Байесовский вывод для монетки, сопряжённые априорные распределения



Машинное обучение
Основы байесовского вывода

02

Байесовский вывод для монетки

Random facts:

- 11 сентября — не самая приятная дата: в 1847 г. 300 человек погибли в результате урагана на Ньюфаундленде, в 1927 г. произошло землетрясение в Крыму с большими разрушениями от Севастополя до Феодосии, в 2012 г. в пожарах на двух ткацких фабриках в Пакистане погибли 315 сотрудников, в 2015 г. башенный кран упал на мечеть Аль-Харам в Мекке и задавил насмерть 111 человек, а в 2001 году — сами понимаете...
- 11 сентября — Национальный день Каталонии, учреждённый в память о взятии Барселоны во время войны за испанское наследство 11 сентября 1713 года



Байесовский вывод для монетки

- 11 сентября 1943 г. на стекольном заводе в Гусь-Хрустальном был выпущен первый советский гранёный стакан; Вера Мухина здесь была ни при чём
- 11 сентября 1962 г. The Beatles завершили запись своего первого сингла P.S. I Love You; на барабанах был Энди Уайт, сессионный ударник, которого пригласила студия, потому что Ринго очень волновался и никак не мог выдать нужный результат
- 11 сентября 1973 г. в результате профинансированного ЦРУ переворота в Чили президент Альенде совершил самоубийство, а к власти пришёл Аугусто Пиночет; так прекратился проект Киберсин о централизованном компьютерном управлении плановой экономикой под руководством Стаффорда Бира

ML vs. MAP

- Мы остановились на том, что в статистике обычно ищут **гипотезу максимального правдоподобия** (maximum likelihood):

$$\theta_{ML} = \arg \max_{\theta} p(D | \theta).$$

- В байесовском подходе ищут **апостериорное распределение** (posterior)

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

и, возможно, **максимальную апостериорную гипотезу** (maximum a posteriori):

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | D) = \arg \max_{\theta} p(D | \theta)p(\theta).$$

Постановка задачи

- Простая задача вывода: дана нечестная монетка, она подброшена N раз, имеется последовательность результатов падения монетки. Надо определить её «нечестность» и предсказать, чем она выпадет в следующий раз.

Постановка задачи

- Если у нас есть вероятность $p_h = \theta$ того, что монетка выпадет орлом (вероятность решки $p_t = 1 - \theta$), то вероятность того, что выпадет последовательность s , которая содержит n орлов и m решек, равна

$$p(s \mid p_h = \theta) = \theta^n (1 - \theta)^m.$$

- Сделаем предположение: будем считать, что вероятность монетки до экспериментов для нас распределена равномерно, т.е. у нас нет априорных предпочтений насчёт θ .
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

Пример применения теоремы Байеса

- Правдоподобие: $p(\theta | s) = \frac{p(s|\theta)p(\theta)}{p(s)}$.
- Здесь $p(\theta)$ следует понимать как непрерывную случайную величину, сосредоточенную на интервале $[0, 1]$, коей она и является. Наше предположение о равномерном распределении в данном случае значит, что априорная вероятность $p(\theta) = 1, \theta \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s | \theta)$ мы уже знаем.
- Итого получается:

$$p(\theta | s) = \frac{\theta^n (1 - \theta)^m}{p(s)}.$$

Пример применения теоремы Байеса

- Итого получается:

$$p(\theta | s) = \frac{\theta^n (1 - \theta)^m}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$p(s) = \int_0^1 \theta^n (1 - \theta)^m d\theta = \frac{\Gamma(n + 1)\Gamma(m + 1)}{\Gamma(n + m + 2)} = \frac{n!m!}{(n + m + 1)!},$$

но найти $\arg \max_{\theta} p(\theta | s) = \frac{n}{n+m}$ можно и без этого.

Пример применения теоремы Байеса

- Итого получается:

$$p(\theta | s) = \frac{\theta^n (1 - \theta)^m}{p(s)}.$$

- Но это ещё не всё. Чтобы предсказать следующий исход, надо найти $p(\text{heads} | s)$:

$$p(\text{heads}|s) = \int_0^1 p(\text{heads}|\theta)p(\theta|s)d\theta = \int_0^1 \frac{\theta^{n+1}(1-\theta)^m}{p(s)} dp_h = \frac{(n+1)!m!}{(n+m+2)!} \cdot \frac{(n+m+1)!}{n!m!} = \frac{n+1}{n+m+2}.$$

- Получили правило Лапласа.

Пример применения теоремы Байеса

- Итого получается:

$$p(\theta | s) = \frac{\theta^n (1 - \theta)^m}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:

(1) найти апостериорное распределение (posterior distribution) на гипотезах/параметрах:

$$p(\theta | D) \propto p(D|\theta)p(\theta)$$

и/или найти максимальную апостериорную гипотезу (maximum a posteriori)

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | D);$$

(2) найти предсказательное распределение (predictive distribution) исходов дальнейших экспериментов:

$$p(x | D) \propto \int_{\theta \in \Theta} p(x | \theta) p(D|\theta) p(\theta) d\theta.$$

Напоминание

- Напоминаю, что основная наша задача — как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
 - $p(x | \theta)$ — правдоподобие данных;
 - $p(\theta)$ — априорное распределение;
 - $p(x) = \int_{\Theta} p(x | \theta)p(\theta)d\theta$ — маргинальное правдоподобие;
 - $p(\theta | x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ — апостериорное распределение;
 - $p(x' | x) = \int_{\Theta} p(x' | \theta)p(\theta | x)d\theta$ — предсказание нового x' .
- Задача обычно в том, чтобы найти $p(\theta | x)$ и/или $p(x' | x)$.

Априорные распределения

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё **априорное распределение** (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода — как подсчитать $p(\theta \mid x)$ и/или $p(x' \mid x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать априорные распределения?

Сопряжённые априорные распределения

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и **a posteriori**.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta | x)$.
- Я хочу, чтобы $p(\theta | x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

Сопряжённые априорные распределения

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством **сопряжённых априорных распределений** для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются **гиперпараметрами** (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

Сопряжённые априорные распределения

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

Испытания Бернулли

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Испытания Бернулли

- Плотность распределения нечестности монетки θ

$$p(\theta | \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f | \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$p(\theta | s, f) = \frac{\binom{s+f}{s} \theta^{s+\alpha-1} (1-\theta)^{f+\beta-1} / B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s} x^{s+\alpha-1} (1-x)^{f+\beta-1} / B(\alpha, \beta) dx} = \frac{\theta^{s+\alpha-1} (1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}.$$

Испытания Бернулли

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ — это

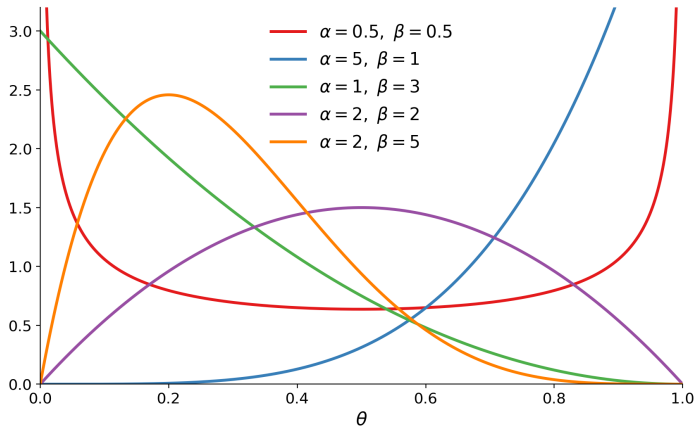
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s + \alpha, f + \beta) \propto \theta^{s+\alpha-1} (1 - \theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

Бета-распределение



Мультиномиальное распределение

- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i .
- Параметры θ_i показывают вероятность попасть в категорию i :

$$p(x \mid \theta) = \binom{n}{x_1, \dots, x_n} \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

Мультиномиальное распределение

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

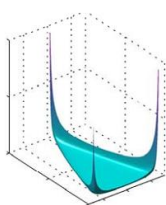
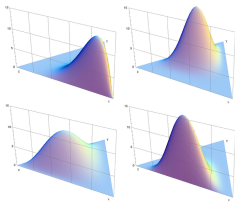


Домашнее
изменятся на

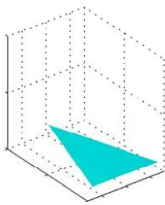
задание. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1+\alpha_1-1} \theta_2^{x_2+\alpha_2-1} \dots \theta_k^{x_k+\alpha_k-1}.$$

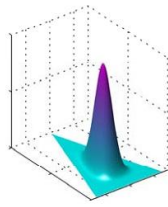
Распределение Дирихле



$$\{\alpha_k\} = 0.1$$



$$\{\alpha_k\} = 1$$



$$\{\alpha_k\} = 10$$

О болезнях и вероятностях

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?

О болезнях и вероятностях

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?
- Ответ: 16%.

Доказательство

- Обозначим через t результат теста, через d — наличие болезни.
- $p(t = 1) = p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)$.
- Используем теорему Байеса:

$$p(d = 1|t = 1) = \frac{p(t = 1|d = 1)p(d = 1)}{p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)} = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16.$$

Вывод

- Вот такие задачи составляют суть вероятностного вывода (probabilistic inference).
- Поскольку они обычно основаны на теореме Байеса, вывод часто называют байесовским (Bayesian inference).
- Но не только поэтому.

Упражнения

1. У моего знакомого два ребёнка. Будем предполагать, что пол ребёнка выбирается независимо и равновероятно, с вероятностью $\frac{1}{2}$. Две постановки вопроса:
 - (1) я спросил, есть ли у него мальчики, и он ответил «да»; какова вероятность того, что один из детей — девочка?
 - (2) я встретил одного из его детей, и это мальчик; какова вероятность того, что второй ребёнок — девочка?

Упражнения

2. Произошло убийство. На месте убийства найдена кровь, которая явно принадлежит убийце. Кровь принадлежит редкой группе, которая присутствует у 1% населения, в том числе у подсудимого.
 - (1) Прокурор говорит: «Шанс, что у подсудимого была бы именно такая группа крови, если бы он был невиновен — всего 1%; значит, с вероятностью 99% он виновен». В чём не прав прокурор?
 - (2) Адвокат говорит: «В городе живёт миллион человек, то есть у 10000 из них такая группа крови. Значит, всё, что говорит нам эта кровь — это что подсудимый совершил убийство с вероятностью 0.01%; никакое это не доказательство». В чём не прав адвокат?

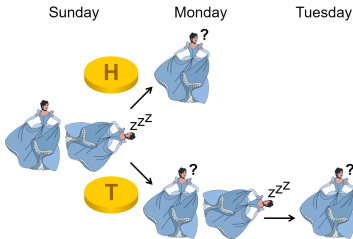
Упражнения

3. Реальные случаи.

- (1) Прокурор указал, что O.J. Simpson уже бил жену в прошлом. Адвокат ответил: «Убивают только одну из 2500 женщин, подвергавшихся семейному насилию, так что это вообще нерелевантно». Суд согласился с адвокатом; верно ли это рассуждение?
- (2) У Sally Clark погибли два младенца; прокурор указал, что вероятность двух случаев SIDS в одной семье, которую он получил из статистики одиночных случаев, — около 1 из 73 миллионов; в чём он не прав?

Задача о спящей красавице

- А вот парадокс, который до сих пор вызывает споры (Zuboff, 1990; Elga, 2000). В воскресенье Спящую красавицу усыпляют и бросают монетку:
 - если выпадает орёл, её будят в понедельник и опять усыпляют до среды
 - если решка, то будят в понедельник, делают укол, от которого она забывает о пробуждении, и будят ещё раз во вторник, а потом так же в среду



Задача о спящей красавице

- Спящая красавица знает всю процедуру эксперимента, но не может различить разные пробуждения и отличить понедельник от вторника, когда просыпается
- Представьте себя на месте Спящей красавицы; вас разбудили; какова вероятность того, что монета упала орлом?

