

Введение в байесовский вывод

Сергей Николенко

Академия MADE — Mail.Ru

17 апреля 2020 г.

Random facts:

- 17 апреля 1492 г. был утверждён один из важнейших научных грантов: Христофор Колумб подписал контракт с Испанией, обязуясь открыть новый путь в Индию
- 17 апреля 1521 г. Мартин Лютер был отлучён от лона Римской Католической Церкви
- 17 апреля 1607 г. католическая церковь взяла реванш: 21-летний Арман Жан дю Плесси де Ришелье был посвящён в сан епископа
- 17 апреля 1722 г. Пётр I ввёл подать на ношение бороды: 50 рублей в год!
- 17 апреля 1986 г. постановление ЦК КПСС «Об основных направлениях ускорения решения жилищной проблемы в стране» пообещало, что у каждой семьи к 2000 году будет отдельная квартира или дом
- 17 апреля 2011 г. на телеканале HBO вышла первая серия *Game of Thrones*

Что такое машинное обучение

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

- 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...

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

- Термин AI и формулировки основных задач появились 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.

Мы предлагаем исследование искусственного интеллекта сроком в 2 месяца с участием 10 человек летом 1956 года в Дартмутском колледже, ГанOVER, 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.

- 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.

- 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).

- 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

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

- 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.

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

Определение

- Что значит — обучающаяся машина? Как определить «обучаемость»?
- Определение из книги Митчелла: «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

- *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.

- По сути машинное обучение – это наука о неопределённости: мы пытаемся вывести значения параметров из неполных данных, порождённых с шумом/ошибками, это заведомо статистическая задача.
- Кроме того, 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.

- 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.

Байесовский подход

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

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

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%.

- 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:

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

- 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.

- 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.

- Мы можем рассуждать о том, «насколько вероятно» то, что
 - сборная России победит на чемпионате мира по футболу в 2022 году;
 - «Одиссею» написала женщина;
 - Керенский бежал за границу в женском платье;
 - ...
- Но о «стремящемся к бесконечности количестве экспериментов» говорить бессмысленно — эксперимент здесь ровно один.

- 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).

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.

- 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).
- Вообще, *функция правдоподобия* имеет вид

$$a \mapsto p(y|x = a)$$

для некоторой случайной величины y .

- 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.
- Гипотеза максимального правдоподобия скажет, что вероятность решки равна числу выпавших решек, делённому на число экспериментов.

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.
- Гипотеза максимального правдоподобия скажет, что вероятность решки равна числу выпавших решек, делённому на число экспериментов.
- То есть если вы взяли незнакомую монетку, подбросили её один раз и она выпала решкой, вы теперь ожидаете, что она всегда будет выпадать только решкой, правильно?
- Странно получается... давайте поговорим об этом поподробнее.

Байесовский вывод для монетки

- Итак, 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 того, что монетка выпадет решкой (вероятность орла $p_t = 1 - p_h$), то вероятность того, что выпадет последовательность s , которая содержит n_h решек и n_t орлов, равна

$$p(s|p_h) = p_h^{n_h} (1 - p_h)^{n_t}.$$

- Сделаем предположение: будем считать, что монетка выпадает равномерно, т.е. у нас нет априорного знания p_h .
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

Пример применения теоремы Байеса

- Правдоподобие: $p(s|p_h) = \frac{p(s|p_h)p(p_h)}{p(s)}$.
- Здесь $p(p_h)$ 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 $p(p_h) = 1, p_h \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|p_h)$ мы уже знаем.
- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1 - p_h)^{n_t}}{p(s)}.$$

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$\begin{aligned} p(s) &= \int_0^1 p_h^{n_h}(1-p_h)^{n_t} dp_h = \\ &= \frac{\Gamma(n_h+1)\Gamma(n_t+1)}{\Gamma(n_h+n_t+2)} = \frac{n_h!n_t!}{(n_h+n_t+1)!}, \end{aligned}$$

но найти $\arg \max_{p_h} p(p_h | s) = \frac{n_h}{n_h+n_t}$ можно и без этого.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

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

$$\begin{aligned} p(\text{heads}|s) &= \int_0^1 p(\text{heads}|p_h)p(p_h|s)dp_h = \\ &= \int_0^1 \frac{p_h^{n_h+1}(1-p_h)^{n_t}}{p(s)} dp_h = \\ &= \frac{(n_h+1)!n_t!}{(n_h+n_t+2)!} \cdot \frac{(n_h+n_t+1)!}{n_h!n_t!} = \frac{n_h+1}{n_h+n_t+2}. \end{aligned}$$

- Получили правило Лапласа.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:
 1. найти апостериорное распределение на гипотезах/параметрах:

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

(и/или найти гипотезу максимального правдоподобия $\arg \max_{\theta} p(\theta | D)$);

2. найти апостериорное распределение исходов дальнейших экспериментов:

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

Сопряжённые априорные распределения

- 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 $p(x' | x)$.

- 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 $p(\theta | x)$ и/или $p(x' | x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать 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 $p(\theta | \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x | \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta | x, \alpha)$ остаётся в том же семействе: $p(\theta | x, \alpha) = p(\theta | \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

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

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

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

- 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

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

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

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

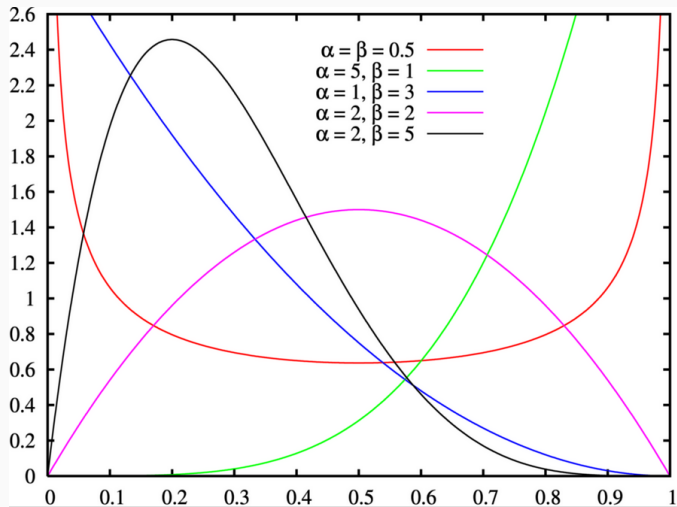
$$p(\theta | \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}.$$

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

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

- 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 :

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

- Сопряжённым априорным распределением будет распределение Дирихле:

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

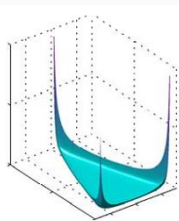
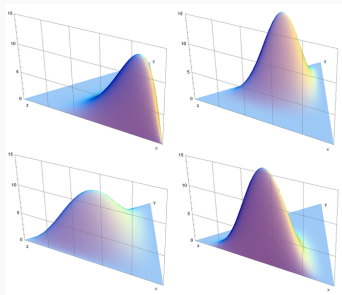
- Сопряжённым априорным распределением будет распределение Дирихле:

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

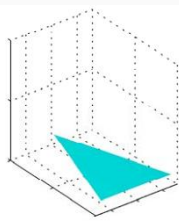
Упражнение. Докажите, что при получении данных x_1, \dots, x_k гиперпараметры изменятся на

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

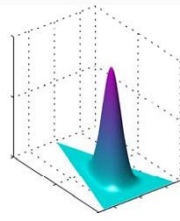
Распределение Дирихле



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



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



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

Спасибо!

Спасибо за внимание!