ПРИНЦИП МАКСИМУМА ЭНТРОПИИ

Сергей Николенко

СПбГУ — Санкт-Петербург

21 ноября 2024 г.

Random facts:

- 21 ноября 1620 г. после трёхмесячного плавания 102 переселенца (не считая детей) сошли с борта «Мэйфлауэра» в районе мыса Код; только 37 из них были пуританами-диссентерами
- 21 ноября 1783 г. состоялся первый в истории полёт людей на воздушном шаре, а 21 ноября 1877 г. Томас Эдисон изобрёл фонограф
- 21 ноября 1916 г. затонул «Британник» — корабль-близнец «Олимпика» и «Титаника»; он так и не успел стать пассажирским лайнером, в ноябре 1915 г. был превращён в госпитальное судно, а через год налетел на немецкую мину в Средиземном море; на «Британнике» было около 40 спасательных шлюпок, а температура воды была выше 20°C, так что при крушении почти никто не погиб
- 21 ноября 2012 г. Конгресс США официально отменил поправку Джексона—Вэника, принятая против Советского Союза в 1974 году
- 21 ноября 2013 г. в центре Киева начался Евромайдан; уже через год новый президент Украины Пётр Порошенко объявил 21 ноября национальным праздником как День достоинства и свободы

Пример Джейнса

Пример Джейнса

- Рассмотрим простую задачу байесовского вывода: предположим, что перед нами урна, в которой N шаров, R из которых красные, а $N - R$ белые
- Мы достаём из урны n шаров, r из которых оказываются красными
- Легко найти распределение пар (n, r) по заданным N и R: вероятность вынуть из урны сначала r красных шаров подряд, а затем w белых, составляет

$$\frac{R}{N}\frac{R-1}{N-1}\cdot\ldots\cdot\frac{R-r+1}{N-r+1}\cdot\frac{N-R}{N-r}\frac{N-R-1}{N-r-1}\cdot\ldots\cdot\frac{N-R-w+1}{N-r-w+1} =$$
$$= \frac{R!(N-r)!}{(R-r)!N!}\cdot\frac{(N-R)!(N-r-w)!}{(N-R-w)!(N-r)!} = \frac{R!(N-R)!(N-n)!}{(R-r)!(N-R-w)!N!}.$$

Пример Джейнса

- Эта вероятность получена для конкретной последовательности исходов, и она по симметрии не зависит от последовательности, то есть достаточно умножить её на число возможных таких последовательностей $\binom{n}{r}$:

$$p\left(r|N,R,n\right) = \frac{R!(N-R)!(N-n)!}{(R-r)!(N-R-w)!N!}\frac{n!}{r!(n-r)!} = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}},$$

  то есть гипергеометрическое распределение

Пример Джейнса

- Но теперь перед нами стоит обратная задача: мы получили данные $D = (n, r)$ и хотим что-то сказать о значениях N и R
- Байесовский вывод проходит точно так же, как выше:

$$p\left(N, R|D\right) = \frac{p\left(D|N, R\right) p\left(R|N\right) p\left(N\right)}{p\left(D\right)},$$

и здесь в знаменателе стоит нормировочная константа

$$p\left(D\right) = \sum_{N=0}^{\infty} \sum_{R=0}^{N} p\left(D|N, R\right) p\left(R|N\right) p\left(N\right),$$

а правдоподобие представляет собой гипергеометрическое распределение

Пример Джейнса

- Самый интересный вопрос: как выбрать априорное распределение $p(N, R)$? Мы его уже разложили в произведение

$$p(N, R) = p(R|N)\, p(N);$$

  в байесовском выводе мы выбираем априорное распределение как захотим, согласно своей интуиции, но чего мы можем захотеть от этого априорного распределения, как формализовать интуицию?
- Во-первых, есть интуиция о том, что апостериорное распределение на величину N должно зависеть от числа выбранных шаров n (как в случае танков), но не должно зависеть от того, сколько из них оказались красными: с какой стати пропорция красных шаров должна влиять на их общее число?

Пример Джейнса

- Более того в данном случае хотелось бы, чтобы полученные данные «обрезали» невозможные значения, то есть сообщали нам, что $N \geq n$, но не меняли нашу априорную интуицию о том, сколько именно шаров может быть в урне, если это число больше n. Иными словами, хотелось бы получить

$$p\left(N|D\right) = \begin{cases} c \cdot p\left(N\right), & \text{если } N \geq n, \\ 0 & \text{в противном случае,} \end{cases}$$

  где c — нормировочная константа, не зависящая от N
- Это уже нетривиальное требование! Оно накладывает серьёзные ограничения на $p\left(N, R\right)$

Пример Джейнса

- Апостериорное распределение на N равно

$$p\left(N|D\right) = \frac{p\left(N\right)p\left(D|N\right)}{p\left(D\right)} = \frac{p\left(N\right)\sum_{R=0}^{N} p\left(D|N,R\right)p\left(R|N\right)}{p\left(D\right)},$$

  и требование выше значит, что

$$p\left(D|N\right) = \sum_{R=0}^{N} p\left(D|N,R\right)p\left(R|N\right) = \begin{cases} f(n,r), & \text{если } N \geq n, \\ 0 & \text{в противном случае,} \end{cases}$$

  где $f(n,r)$ — это некоторая функция, зависящая от данных n и r, но не от N

Пример Джейнса

- Более того, можно вспомнить, что мы говорим о гипергеометрическом распределении: для $N \geq n$

$$\sum_{R=0}^{N} \binom{R}{r} \binom{N-R}{n-r} p\left(R|N\right) = f(n,r) \binom{N}{n}$$

- Это уже весьма нетривиальное условие на априорное распределение $p\left(R|N\right)$!
- Мы получили его из базовой интуиции о том, что данные не должны менять наше мнение об N (это в некотором смысле независимость), но мы вряд ли смогли бы записать такое условие интуитивно, без этих рассуждений
- Разумные априорные распределения на R будут удовлетворять этому условию, и байесовский вывод по N будет тривиальным: апостериорное распределение пропорционально априорному, за исключением начального отрезка $N < n$, который данные делают невозможным

Пример Джейнса

- Осталось выбрать $p\left(R|N\right)$
- Первый вариант — равномерное распределение $p\left(R|N\right) = \frac{1}{N+1}$ на интервале $0, \dots, N$
- Условие легко получить из свойств суммирования биномиальных коэффициентов: мы знаем, что $\sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1}$ (слева выбираем один элемент из $N+1$, фиксируя R, а потом выбираем r из R слева от него и $n-r$ из $N-R$ справа, то есть по сути выбираем $n+1$ элемент из $N+1$), а значит,

$$\sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r} p\left(R|N\right) = \binom{N+1}{n+1}\frac{1}{N+1} =$$

$$= \frac{(N+1)!(N+1-n-1)!}{(n+1)!(N+1)} = \frac{1}{n+1}\binom{N}{n},$$

то есть $f(n,r) = \frac{1}{n+1}$, и эта функция действительно не зависит от N

Пример Джейнса

- Апостериорное распределение тогда получается таким:

$$p\left(R|D,N\right)=\binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r};$$

  уже не гипергеометрическое, потому что по R, а не по N
- Полученное апостериорное распределение удовлетворяет нескольким естественным требованиям «здравого смысла»:
  - для $n=r=0$ это распределение превращается в априорное распределение $p\left(R|N\right)=\frac{1}{N+1}$;
  - для $n=r=1$ (вытащили один шар, и он оказался красным) получается $p\left(R|D=(1,1),N\right)=\frac{2R}{N(N+1)}$, что пропорционально R; это тоже естественно, потому что в правдоподобии вероятность вытащить красный шар $p\left(r=1|n=1,R,N\right)=\frac{R}{N}$, и мы умножаем это правдоподобие на равномерное распределение;
  - кроме того, $\frac{2R}{N(N+1)}$ равно нулю при $R=0$ (что логически невозможно), и это свойство сохраняется для всех $n\geq r\geq 1$

Пример Джейнса

- Кроме того, можно провести байесовский вывод дальше, почти в точности как для монетки
- Для нашего апостериорного распределения

$$p\left(R|D,N\right)=\binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r}$$

максимальная апостериорная гипотеза будет (проверьте в качестве упражнения)

$$R_{\mathrm{MAP}}=\left\lfloor\frac{r}{n}\left(N+1\right)\right\rfloor,$$

что тоже звучит вполне естественно

Пример Джейнса

- Но этот максимум не совпадает с математическим ожиданием:

$$\mathbb{E}[R] = \sum_{R=0}^{N} R{\cdot}p(R|D,N) = \binom{N+1}{n+1}^{-1} \sum_{R=0}^{N} R\binom{R}{r}\binom{N-R}{n-r} =$$

$$= \binom{N+1}{n+1}^{-1} \left(\sum_{R=0}^{N} (R+1)\binom{R}{r}\binom{N-R}{n-r} - \sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r} \right) =$$

$$= \binom{N+1}{n+1}^{-1} \left(\sum_{R=0}^{N} (r+1)\binom{R+1}{r+1}\binom{N-R}{n-r} - \sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r} \right) =$$

$$= (r+1)\binom{N+1}{n+1}^{-1}\binom{N+2}{n+2} - 1 = \frac{(N+2)(r+1)}{n+2} - 1,$$

где мы воспользовались известным тождеством о биномиальных коэффициентах: $(R+1)\binom{R}{r} = (r+1)\binom{R+1}{r+1}$

Пример Джейнса

- Для больших N, r и n этот результат, конечно, похож на R_{MAP}, но поправка существенна: ожидаемая доля красных шаров, остающихся в урне, составляет

$$\frac{\mathbb{E}[R] - r}{N - n} = \frac{(N+2)(r+1) - (n+2) - r(n+2)}{(N-n)(n+2)} = \frac{r+1}{n+2},$$

  то есть мы опять получили в точности правило Лапласа

Пример Джейнса

- Можно и предсказательное распределение в виде правила Лапласа получить (опять упражнение):

$$p\left(\mathrm{Red}|D,N\right)=\sum_{R=0}^{N}R\cdot p\left(\mathrm{Red},R|D,N\right)=$$
$$=\sum_{R=0}^{N}R\cdot p\left(\mathrm{Red}|R,D,N\right)p\left(R|D,N\right)=$$
$$=\binom{N+1}{n+1}^{-1}\sum_{R=0}^{N}\frac{R-r}{N-n}\binom{R}{r}\binom{N-R}{n-r}=\frac{r+1}{n+2}.$$

Пример Джейнса

- Аналогично (и снова упражнение) можно найти дисперсию апостериорного распределения:

$$\hat{R} = r + (N-n)p \pm \sqrt{\frac{p(1-p)}{n+3}(N+2)(N-n)},$$

  где мы обозначили $p = \frac{r+1}{n+2}$, а также подсчитали дисперсию именно оценки R, а не оценки вероятности $\frac{R-r}{N-n}$
- Оценка вероятности становится точнее с ростом n, то есть с ростом полученной выборки, причём падает до нуля, если $n = N$ — это очень интуитивно!

Пример Джейнса

- А если вернуться к оценке доли шаров, то нам нужно будет разделить предыдущую оценку на $N - n$, и мы получим

$$\frac{\hat{R} - r}{N - n} = p \pm \sqrt{\frac{p(1-p)}{n+3} \frac{N+2}{N-n}}$$

- Теперь ситуация обратная: чем больше шаров мы вынули из урны, тем менее точной будет оценка доли оставшихся, а в пределе $N \to \infty$ останется как раз $p \pm \sqrt{\frac{p(1-p)}{n+3}}$, что в точности соответствует дисперсии монетки
- Но давайте исследуем и другие формы априорных распределений!

Пример Джейнса

- Сначала сделаем совсем небольшое изменение: предположим, что нам заранее известно, что в урне есть хотя бы один красный и хотя бы один белый шар, а в остальном оставим распределение равномерным: $p\left(R|N\right) = \frac{1}{N-1}$ на интервале $1, \dots, N-1$
- Тогда нужно выбросить слагаемые, соответствующие $R = 0$ и $R = N$; для $R = 0$ $\binom{R}{r} = [r = 0]$, а для $R = N$ $\binom{N-R}{n-r} = [r = n]$, и наше тождество превращается в

$$\sum_{R=1}^{N-1} \binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1} - \binom{N}{n}[r = n] - \binom{N}{n}[r = 0]$$

- Соответственно, апостериорное распределение теперь

$$p\left(R|D,N\right) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N+1}{n+1} - \binom{N}{n}[r = n] - \binom{N}{n}[r = 0]}$$

Пример Джейнса

- Первое наблюдение: если мы получили данные с $0 < r < n$ (среди выбранных были и красные, и белые шары), то вся разница пропадает, и апостериорное распределение остаётся тем же
- Это кажется контринтуитивным: мы изменили вероятностное пространство, а распределение осталось тем же самым! Джейнс пишет так: «И всё же после некоторого размышления мы видим, что результат верен, поскольку в этом случае из данных можно дедуктивно умозаключить, что R не может быть равным нулю или N; поэтому не важно, говорит ли нам то же самое ещё и априорное распределение: состояние наших знаний об R в этих двух случаях одно и то же, на что указывают и теория вероятностей, и логика»

Пример Джейнса

- А, например, для $r = 0$, когда мы не вытащили ни одного красного шара, результат изменится:

$$p\left(R|D = (0, n), N\right) = \binom{N-R}{n}\left(\binom{N+1}{n+1} - \binom{N}{n}\right)^{-1} = \\ = \binom{N}{n+1}\binom{N-R}{n}^{-1}$$

  для $R = 1, \dots, N-1$ и 0 вне этого интервала
- Это в точности совпадает с апостеориорным распределением выше, только перенормированным на константу $\frac{N+1}{N-n}$
- Хотя на этот раз изменение в априорном распределении исключило аж максимальную апостериорную гипотезу $R = 0$, общий вид распределения на оставшейся части носителя всё равно не изменился; разные априорные распределения не обязательно приводят к разным выводам

Пример Джейнса

- Идём дальше: правило Лапласа говорит, что предсказание при равномерном распределении сглаживается, причём это сглаживание в точности соответствует правдоподобию данных, в которых мы достали из урны два шара, и один из них оказался красным (эквивалентный размер выборки)
- Иначе говоря, равномерное априорное распределение даёт при выводе некоторую информацию, соответствующую выборке из $n = 2$ шаров
- Можно ли придумать полностью неинформативное (uninformative) априорное распределение, такое, чтобы никакой добавки не получалось?

Пример Джейнса

- Давайте искать это распределение так: с какого распределения нужно начать, чтобы после получения данных $D = (n, r) = (2, 1)$ прийти к равномерному?
- Иначе говоря, на какое априорное распределение нужно умножить правдоподобие

$$p\left(r = 1 | N, R, n = 2\right) = \frac{\binom{R}{1}\binom{N-R}{1}}{\binom{N}{2}} = \frac{R(N-R)}{N(N-1)},$$

  чтобы на выходе получилась константа?
- На первый взгляд кажется, что ничего не выйдет, ведь для $R = 0$ и $R = N$ получается $p\left(r = 1 | N, R, n = 2\right) = 0$

Пример Джейнса

- Но мы только что придумали, как с этим справиться: отрежем крайние случаи в априорном распределении,
  и апостериорное не изменится, если $1 \leq r \leq n-1$, а у нас данные именно такие!
- Таким образом, для априорного распределения

$$p\left(R|N\right) = \frac{\text{const}}{R(N-R)} \quad \text{для} \quad R = 1, \ldots, N-1,$$

апостериорное распределение будет равно

$$p\left(R|D=(n,r),N\right) = \frac{\text{const}}{R(N-R)} \binom{R}{r} \binom{N-R}{n-r} = \\ = \frac{\text{const}}{r(n-r)} \binom{R-1}{r-1} \binom{N-R-1}{n-r-1},$$

Пример Джейнса

- А из этого по формуле суммирования можно и нормировочную константу получить: для $R = 1, \dots, N-1$

$$p\left(R|D=(n,r),N\right) = \binom{N-1}{n-1}^{-1}\binom{R-1}{r-1}\binom{N-R-1}{n-r-1}$$

- Если подставить $n = 2$ и $r = 1$, мы получим равномерное распределение $\frac{1}{N-1}$; можно доказать, что и требуемое свойство выполняется, то есть можно вести вывод отдельно на R, и получить формулы ожидания и дисперсии апостериорного распределения (упражнение)

Пример Джейнса

- Последний случай — «распределение биномиальных обезьян» (binomial monkey prior): предположим, что урна наполнялась шарами, каждый из которых имел некоторую вероятность g быть красным, причём цвет шаров выбирался независимо друг от друга (обезьяны-дальтоники бросают туда случайно выбранные мячи)
- Тогда априорное распределение будет биномиальным:

$$p\left(R|N\right) = \binom{N}{R} g^R (1-g)^{N-R}$$

- Это значит, что априори мы можем оценить число красных шаров в урне как $\hat{R} = Ng \pm \sqrt{Ng(1-g)}$

Пример Джейнса

- Теперь проведём байесовский вывод; проверим наше свойство:

$$p\left(D|N\right)=\sum_{R=0}^{N}p\left(D|N,R\right)p\left(R|N\right)=\sum_{R=0}^{N}\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}\binom{N}{R}g^{R}(1-g)^{N-R}=$$

$$=\binom{n}{r}\sum_{R=0}^{N}\binom{N-n}{R-r}g^{R}(1-g)^{N-R}=$$

$$=\binom{n}{r}g^{r}(1-g)^{n-r}\sum_{R=0}^{N}\binom{N-n}{R-r}g^{R-r}(1-g)^{N-n-R}=\binom{n}{r}g^{r}(1-g)^{n-r},$$

где мы воспользовались тем, что (проверьте!)

$$\binom{R}{r}\binom{N-R}{n-r}\binom{N}{R}=\binom{N}{n}\binom{n}{r}\binom{N-n}{R-r},$$

а затем просуммировали по биномиальному распределению, выбирающему $R-r$ красных шаров из $N-n$ попыток; результат не зависит от N, так что свойство выполняется

Пример Джейнса

- Подсчитаем апостериорное распределение:

$$p\left(R|D,N\right) = \text{const} \cdot \binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r} \binom{N-r}{n-r},$$

и нормировочную константу можно подсчитать как

$$1 = \sum_{R=0}^{N} p\left(R|D,N\right) = \text{const} \binom{N}{n} \binom{n}{r} \sum_{R=0}^{N} \binom{N-n}{R-r} g^R (1-g)^{N-R} =$$

$$= \text{const} \binom{N}{n} \binom{n}{r} g^r (1-g)^{n-r}$$

для $R = r, \ldots, N-n+r$

Пример Джейнса

- Теперь можно выразить const и найти апостериорное распределение уже полностью, вместе с константой:

$$p\left(R|D,N\right) = \binom{N-n}{R-r} g^{R-r}(1-g)^{N-R-n+r}$$

- Отсюда можно найти и среднее, и дисперсию апостериорной оценки $\hat{R}$:

$$\hat{R} = r + (N-n)g \pm \sqrt{g(1-g)(N-n)},$$

и мы получаем оценку на долю оставшихся красных шаров

$$\frac{\hat{R}-r}{N-n} = g \pm \sqrt{\frac{g(1-g)}{N-n}}.$$

Пример Джейнса

- Раньше апостериорное распределение на p определялось в первую очередь данными $D = (n, r)$, а теперь данные вообще практически не участвуют в оценках!
- Значение r для оценки теперь вообще не важно, а значение n просто постепенно уменьшает дисперсию
- Биномиальное априорное распределение *отменяет влияние данных*: апостериорная оценка совпадает с априорной и никак не зависит от r и n!
- Это выглядит странно, но, опять же, вполне интуитивно: априорное распределение с «биномиальными обезьянами» предполагало, что цвет шаров выбирается независимо друг от друга — вот и получилось, что знание цвета части шаров ничего не сообщает нам о цвете остальных
- Эта независимость оказалась «пропущена через» байесовский вывод и сохранилась в апостериорном результате

Пример Джейнса

- Итак, мы рассмотрели пример об урне, из которой достают красные и белые шары, и провели байесовский вывод для нескольких разных априорных распределений:
  - обычного равномерного распределения на $R = 0, \dots, N$;
  - обрезанного равномерного, в котором исключены крайние случаи только красных и только белых шаров в урне, то есть равномерного распределения на $R = 1, \dots, N-1$;
  - неинформативного распределения $\frac{\text{const}}{R(N-R)}$, которое не даёт дополнительного сглаживания в апостериорном распределении;
  - распределения биномиальных обезьян $p\left(R|N\right) = \binom{N}{R} g^R (1-g)^{N-R}$, которое, как выяснилось, полностью отменяет влияние данных и сохраняет независимость цвета шаров в урне друг от друга.
- Тем самым мы рассмотрели несколько важных частных случаев, которые дают представление о том, насколько разных и интересных эффектов можно добиться выбором априорного распределения; пока на дискретном примере...

Принцип максимума энтропии

Принцип максимума энтропии

- Обычно энтропию выводят из аксиомы непротиворечивости

$$H_3(p_1, p_2, p_3) = H_2(p_1, p_2 + p_3) + qH_2\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right)$$

- Но давайте посмотрим на понятие энтропии с другой стороны
- Пусть у нас имеется некоторая информация I, которой должно соответствовать распределение вероятностей $(p_1, \dots, p_n)$
- "Информация"" здесь понимается в самом общем виде: это просто значит, что не любое распределение нам подойдёт

Принцип максимума энтропии

- Будем порождать распределения так: разобьём общую вероятность (единицу) на N маленьких частей (квантов) размера $\delta = \frac{1}{N}$ и будем случайно распределять их между n исходами: N раз бросать честную игральную кость с n гранями
- Тогда, если $p_i = \frac{N_i}{N}$, нам нужно будет распределить ровно N_i этих квантов в i-й исход
- Каждая последовательность из N бросаний выпадает с вероятностью n^{-N}, и вероятность получить распределение $(p_1, \dots, p_n)$ получится из числа сочетаний:

$$n^{-N} \frac{N!}{N_1! \cdot \dots \cdot N_n!}$$

Принцип максимума энтропии

- Если мы будем запускать этот процесс и выбрасывать распределения, не соответствующие I, какое распределение будет чаще всего попадаться?
- Это будет то из распределений, соответствующих I, которое максимизирует величину $\frac{N!}{N_1!\cdot\ldots\cdot N_n!}$; для $N \to \infty$

$$\log N! = N\log N - N + \sqrt{2\pi N} + \frac{1}{12N} + O\left(\frac{1}{N^2}\right)$$

- Поскольку $N_i = p_i N$ и $\sum_{i=1}^{n} N_i = N$, $\log \frac{N!}{N_1!\cdot\ldots\cdot N_n!}$ равен

$$\log N! - \sum_{i=1}^{n} \log N_i! = N\log N - N - \sum_{i=1}^{n} (N_i \log N_i - N_i) + O\left(\sqrt{N}\right) =$$

$$= N\log N - N - \sum_{i=1}^{n} p_i N(\log N + \log p_i) + \sum_{i=1}^{n} N_i + O\left(\sqrt{N}\right) =$$

$$= -N\sum_{i=1}^{n} p_i \log p_i + O\left(\sqrt{N}\right).$$

Принцип максимума энтропии

- Получили

$$\log \frac{N!}{N_1! \cdot \ldots \cdot N_n!} = -N \sum_{i=1}^{n} p_i \log p_i + O\left(\sqrt{N}\right)$$

- При $N \to \infty$ логарифм числа сочетаний растёт как N умножить на энтропию, то есть из такого (вполне естественного) порождающего процесса с наибольшей вероятностью будет получаться именно распределение, максимизирующее энтропию!
- Мы получили другое обоснование понятия энтропии, не опирающееся на аксиому непротиворечивости
- Более того, в этом подходе ещё и вполне естественно получается, что энтропию нужно именно *максимизировать*

Принцип максимума энтропии

- *Принцип максимума энтропии* (maximum entropy principle): для некоторой заданной априорной информации I наиболее характерным распределением вероятностей является распределение с максимальной энтропией среди тех распределений, которые удовлетворяют этой информации
- Рассмотрим кубик, у которого среднее значение составляет $m = 4.5$ вместо 3.5; давайте применим принцип максимума энтропии, то есть максимизируем

$$H\left(p_1, \ldots, p_6\right) = -\sum_{i=1}^{n} p_i \log p_i$$

при двух условиях: $p_1 + \ldots + p_6 = 1$ и $\mathbb{E}\left[i\right] = \sum_{i=1}^{6} i p_i = m$

Принцип максимума энтропии

- Мы можем записать функцию Лагранжа для этой задачи условной оптимизации:

$$L = -\sum_{i=1}^{n} p_i \log p_i - \lambda \sum_{i=1}^{6} i p_i - \mu \sum_{i=1}^{6} p_i,$$

взять от неё производные по p_i и приравнять их к нулю:

$$\frac{\partial L}{\partial p_i} = -\log p_i - 1 - \lambda i - \mu = 0,$$

то есть $p_i = e^{-\lambda i - \mu - 1}$

Принцип максимума энтропии

- Чтобы теперь найти множители Лагранжа λ и μ, нужно применить известные нам условия. Во-первых, $p_1 + ... + p_6 = 1$, то есть

$$\sum_{i=1}^{6} e^{-\lambda i - \mu - 1} = e^{-\mu - 1} \sum_{i=1}^{6} e^{-\lambda i} = 1.$$

- Во-вторых, $\sum_{i=1}^{6} i p_i = m$, то есть

$$\sum_{i=1}^{6} i e^{-\lambda i - \mu - 1} = e^{-\mu - 1} \sum_{i=1}^{6} i e^{-\lambda i} = m.$$

- Отсюда мы можем легко выразить m и μ через λ:

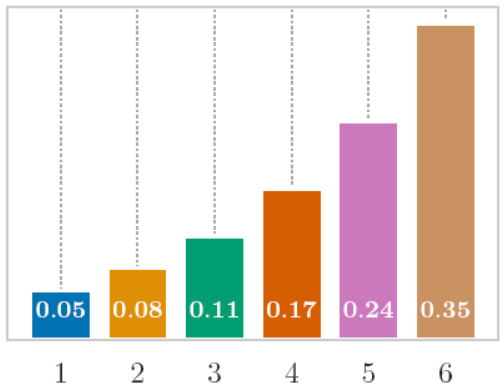
$$m = \frac{\sum_{i=1}^{6} i e^{-\lambda i}}{\sum_{i=1}^{6} e^{-\lambda i}}, \qquad \mu = \log \sum_{i=1}^{6} e^{-\lambda i} - 1$$

Принцип максимума энтропии

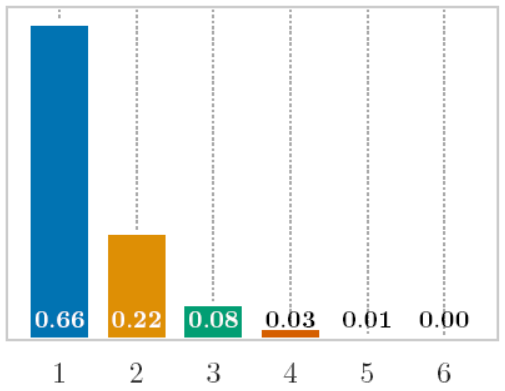
- Теперь осталось найти λ, подставив $m = 4.5$ в первое уравнение (видимо, аналитически не получится), а потом подставить λ во второе и в выражения для p_i
- Для нашего примера численно получается

$$\mathbf{p} = \begin{pmatrix} 0.0544 & 0.0788 & 0.1142 & 0.1654 & 0.2398 & 0.3475 \end{pmatrix}$$

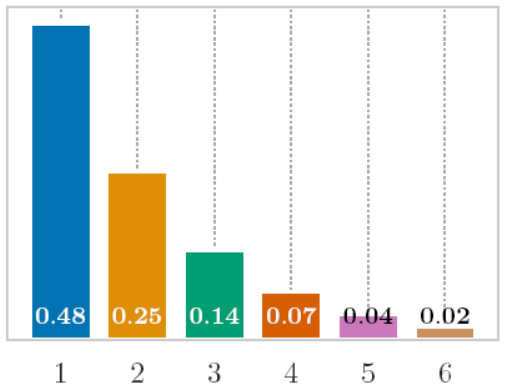
- В сумме они дают единицу, и это распределение скошено в сторону больших значений, как и следовало ожидать из имеющейся априорной информации о среднем



(а) $m = 4.5$



(б) $m = 1.5$



(в) $m = 2$

Принцип максимума энтропии

- Для непрерывного примера давайте попробуем найти распределение на $\mathbb{R}$, достигающее максимальной энтропии с заданным средним μ и дисперсией σ^2
- Мы хотим максимизировать функционал $\int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x$, причём среднее и дисперсия задают нам два условия на эту задачу оптимизации:

$$\int_{-\infty}^{\infty} xp(x) \, \mathrm{d}x = \mu, \qquad \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, \mathrm{d}x = \sigma^2;$$

кроме того, есть ещё базовое условие о том, что $p(x)$ представляет собой плотность вероятности, то есть $\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1$

Принцип максимума энтропии

- Давайте запишем функцию Лагранжа для этой задачи оптимизации, с множителями Лагранжа, соответствующими условиям выше:

$$L(x, \lambda_1, \lambda_2) = \int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x - \lambda_0 \left(1 - \int_{-\infty}^{\infty} p(x) \, \mathrm{d}x \right) -$$
$$- \lambda_1 \left(\mu - \int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x \right) - \lambda_2 \left(\sigma^2 - \int_{-\infty}^{\infty} (x - \mu)^2 p(x) \, \mathrm{d}x \right).$$

Принцип максимума энтропии

- Теперь можно применить метод вариационного исчисления; добавим малую вариацию $\delta p\left(x\right)$ к плотности $p\left(x\right)$ и посмотрим, что произойдёт с функционалом:

$$\delta L = \int_{-\infty}^{\infty} \delta p\left(x\right) \left(\ln p\left(x\right) + 1 + \lambda_0 + \lambda_1 \left(x - \mu\right) + \lambda_2 \left(x - \mu\right)^2\right)$$

- В точке максимума вариация функционала должна быть равна нулю для любого малого $\delta p\left(x\right)$, то есть нулю должно быть равно выражение в скобках, и мы получаем, что

$$0 = \ln p\left(x\right) + 1 + \lambda_0 + \lambda_1 \left(x - \mu\right) + \lambda_2 \left(x - \mu\right)^2,$$

то есть

$$p\left(x\right) = e^{-1-\lambda_0-\lambda_1\left(x-\mu\right)-\lambda_2\left(x-\mu\right)^2}$$

Принцип максимума энтропии

- Получилось, что плотность распределения максимальной энтропии с заданными средним и дисперсией — это экспонента с квадратичной функцией в показателе, то есть нормальное распределение!
- Осталось только подставить полученную плотность в условия, чтобы найти значения λ_0, λ_1 и λ_2 (упражнение); получится, что это гауссиан с именно таким средним и именно такой дисперсией:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

- Метод вариационного исчисления на этом не вполне заканчивается, нужно ещё подтвердить, что результат получился действительно тот, который нам нужен

Принцип максимума энтропии

- Давайте сделаем это другим способом
- Вариационное исчисление дало нам толстый намёк, что ответом должен быть гауссиан; воспользуемся этим намёком и запишем дивергенцию Кульбака–Лейблера между произвольным распределением $q(x)$ и гауссианом $p(x)$:

$$0 \leq \mathrm{KL}(q\|p) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x =$$

$$= -H(q) - \int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x.$$

Принцип максимума энтропии

- Второе слагаемое — перекрёстную энтропию между $q(x)$ и $p(x)$ — можно подсчитать примерно так же, как мы выше делали для двух гауссианов:

$$\begin{aligned}
&\int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x = \\
&\quad = -\int_{-\infty}^{\infty} q(x) \log \sqrt{2\pi\sigma^2} \mathrm{d}x - \int_{-\infty}^{\infty} q(x) \frac{(x-\mu)^2}{2\sigma^2} \mathrm{d}x = \\
&\quad = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} q(x) \frac{(x-\mu)^2}{2\sigma^2} \mathrm{d}x = \\
&\quad = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{\sigma_q^2}{2\sigma^2},
\end{aligned}$$

ведь у нас получилось буквально определение дисперсии распределения $q(x)$

Принцип максимума энтропии

- Но на дисперсию у нас как раз есть условие: $\sigma_q = \sigma$
- А значит,

$$\int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} = -H(p),$$

и теперь дивергенция Кульбака–Лейблера равна

$$0 \leq \mathrm{KL}(q \| p) = -H(q) + H(p), \quad \text{то есть} \quad H(p) \geq H(q)$$

- В итоге мы получили, что для любого распределения $q(x)$ с заданной дисперсией σ^2 его энтропия $H(q)$ не превышает энтропии нормального распределения $H(p)$, что и требовалось
- Это уже полноценное формальное доказательство, и оно вовсе не апеллирует к вариационному исчислению

Спасибо за внимание!