Алгоритмы в биоинформатике, НИУ ВШЭ / Домашняя работа 1

Студент - Катерина Олейникова

Преподаватели - Сергей Спирин, Андрей Миронов

21 Февраля 2022

Примечание Для решения заданий был использован Python. Код приведен в конце отчета.

Задача 1 Вы получили серию результатов испытаний Бернулли:
0 1 0 0 0 0 0 1 0 0
Априорная вероятность распределена по бета-распределению с параметрами $\alpha = 4$, $\beta = 4$.
1. Постройте графики распределения априорной и апостериорной вероятностей.
2. Сделайте MAP и E оценки.

Решение.

Испытания называются испытаниями Бернулли, если каждое из испытаний имеет только два возможных исхода и вероятности исходов остаются неизменными для всех испытаний.
По приведенным в условиях задачи испытаниям Бернулли: число испытаний N = 10, число "успехов"n = 2.

Априорная вероятность распределения какой-то величины есть распределение вероятностей, которое выражает предположения о величине до учета экспериментальных данных.
Апостериорная вероятность – это условная вероятность события при некотором условии, рассматриваемая в противоположность его априорной вероятности.

Бета-распределеение в теории вероятностей и статистике — двухпараметрическое семейство абсолютно непрерывных распределений. Используется для описания случайных величин, значения которых ограничены конечным интервалом.
α и β — параметры априорного распределения (бета-распределения), еще называются гиперпараметрами (в данной задаче равны $\alpha = 4$, $\beta = 4$). Параметры априорного распределения называют гиперпараметрами, чтобы отличить их от параметров модели данных.

- Определим область определения θ для 10 результатов испытаний в интервале от 0 до 1 с помощью функции np.linspace() в Python.

Априорная и апостериорная вероятности связаны через формулу:

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

где P(A|B) - апостериорная вероятность (вероятность гипотезы A при наступлении события B),
P(B|A) - вероятность наступления события B при истинности гипотезы A,
P(A) - априорная вероятность (гипотезы A),
P(B) - полная вероятность наступления события B.

- Определим распределение априорной и апостериорной вероятностей и построим графики через Python.

- Определим MAP и E оценки:
MAP оценка - один из часто используемых методов оценки параметров (максимальная апостериорная

оценка). По формуле (1.11) из книги А.А.Миронова "Биоинформатика последовательностей"можно вычислить MAP оценку в данном случае как

$$\theta = \frac{n + \alpha - 1}{n + m + \alpha + \beta - 2} = .. = 0.3125.$$

Еще одной оценкой параметра является оценка математического ожидания апостериорного распределения, называемая E-оценкой (формула получена из той же книги):

$$E(\theta) = \frac{n + \alpha}{n + m + \alpha + \beta} = .. = 0.333.$$

Мы видим почти такую же оценку, что и для случая MAP-оценки.

Задача 2 Пусть мы имеем некоторое количество наблюдений пуассоновского процесса. Например, это может быть количество прочтений, покрывающих данный ген в нескольких репликах или в нескольких клетках при одноклеточном секвенировании транскриптома. Даже если уровень экспрессии гена постоянен, по случайным причинам мы будем иметь в разных *одинаковых* экспериментах разный уровень покрытия. В простейшем случае можно считать, что число прочтений, покрывающих данный ген, подчиняется распределению Пуассона:

$$Pr(D) = \frac{\lambda^n}{n!} e^{-\lambda}$$

Мы хотим определить уровень экспрессии этого гена λ. Предположим, что λ распределена по Гамма распределению:

$$f(\lambda) = Z^{-1} \lambda^{k-1} exp(-\frac{\lambda}{\theta})$$

с параметрами k = 5, $\theta = 6$. В репликах мы получили нормированные значения количеств прочтений: 8,6,7,12.

1. Построить графики априорного и апостериорного распределений для λ.
2. Сделать MAP и E оценки.

Решение.

Мы хотим определить уровень экспрессии гена λ; напишем формулу для Байесовой оценки параметра:

$$Pr(\lambda|D) = \frac{Pr(D|\lambda) \cdot Pr(\lambda)}{Pr(D)},$$

где $Pr(D|\lambda)$ - произведение вероятностей наблюдений n_i (при этом было сделано M экспериментов и сумарное количество прочтений равно $\sum n_i = N$).

Для данной задачи M = 4, N = 8 + 6 + 7 + 12 = 33.

Теперь зададим формулу для априорного распределения. Уровень экспрессии гена в пространстве всех генов устроен каким-то таким образом:

$$f(x) = \frac{\beta^\alpha}{G(\alpha)} x^{\alpha-1} e^{-\beta x}$$

.

Апостериорное распределение, в свою очередь, станет тоже гамма-распределением с новыми параметрами

$$\alpha' = N + \alpha = (8 + 6 + 7 + 12) + 5 = 38$$

$$\beta' = M + \beta = 4 + 1/6 = 4.167.$$

- Сперва нужно построить априорное и апостериорное распределения для λ (реализовано через Python).
- Сделать MAP и E оценки: для гамма-распределения известны мода и математическое ожидание (см., например, википедию), поэтому можно сразу сделать MAP- и E-оценки:

$$\lambda_{MAP} = \frac{\alpha' - 1}{\beta'} = \frac{38 - 1}{4.167} = 7.923$$

$$\lambda_E = \frac{\alpha'}{\beta'} = \frac{N + \alpha}{M + \beta} = \frac{38}{4.167} = 9.119.$$

Задача 3 Вам дан большой мешок с монетами. В мешке монеты разной кривизны, и есть в том числе монеты с орлами на двух сторонах. Вероятность выпадения орла у односторонней монеты равна 1; вероятность выпадения орла у двусторонней монеты распределена по бета-распределению с параметрами $\alpha = 10$, $\beta = 10$. Распределение вероятностей выпадения орла является смесью Дирихле:

$$f(x) = a \cdot \delta(x-1) + (1-a)Beta_{\alpha,\beta}(x)$$

параметр a распределен по бета-распределению с параметрами $\alpha = 1, \beta = 10$. Схема испытаний: берем монету из мешка, проводим одно испытание и фиксируем результат. Потом берем другую монету и проводим испытание и т.д. В результате получили 80 орлов и 20 решек. Оцените параметр a.

Решение.

Обозначим m = 80 (количество орлов), k = 20 (количество решек), N = 100 (общее количество испытаний).

Параметр a распределен по бета-распределению, тогда:

$$P(a) = \frac{a^{\alpha-1}(1-a)^{\beta-1}}{B(\alpha,\beta)}, (1)$$

где P(a) - априорная вероятность параметра a.

Вероятность выпадения орла у двусторонней монеты = 1/2 (т.к. распределение вероятностей выпадения орла у двусторонней монеты является симметричным); тогда запишем, что

$$P_h = a \cdot 1 + (1-a) \cdot 1/2 = \frac{a+1}{2}$$

$$P_t = a \cdot 0 + (1-a) \cdot 1/2 = \frac{1-a}{2}$$

Формула (1) преобразована в формулу математического ожидания вероятности. В данных формулах первое слагаемое есть случай, когда монета с орлами на двух сторонах (то есть для орла вероятность выпадения - 1, для решки - 0).

Поскольку у нас несколько испытаний, запишем распределение Бернулли:

$$P(D|a) = C_N^m q^m (1-q)^k = C_N^m (\frac{a+1}{2})^m (\frac{1-a}{2})^k = \frac{C_N^m}{2^N}(a+1)^m(1-a)^n$$

Найдем апостериорную вероятность по формуле Байеса:

$$P(a|D) = \frac{P(D|a) \cdot P(a)}{\int P(D|a)P(a)da}$$

$$P(a|D) = \frac{\frac{C_N^m}{2^N}(a+1)^m(1-a)^k a^{\alpha-1}(1-a)^{\beta-1}\frac{1}{B(\alpha,\beta)}}{\int \frac{C_N^m}{2^N}(a+1)^m(1-a)^k a^{\alpha-1}(1-a)^{\beta-1}\frac{1}{B(\alpha,\beta)}da} = \frac{(a+1)^m(1-a)^{k+\beta-1}a^{\alpha-1}}{\int (a+1)^m(1-a)^{k+\beta-1}a^{\alpha-1}da}$$

Подставим входные значения для α, β, k, m:

$$P(a|D) = \frac{(a+1)^{80}(1-a)^{29}}{\int (a+1)^{80}(1-a)^{29}da}$$

После, были построены графики распределения априорной и апостериорной вероятностей значений параметра a в Python.