МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 4** **Москва 2022**

ОБЩИЙ РАЗДЕЛ

УДК 001.102:519.246

Д.В. Виноградов

Утеря научного знания как следствие требований к публикациям

В процессе анализа современных курсов байесовских методов в машинном обучении был обнаружен феномен утери знания — несостоятельности байесовского обучения модели Дирихле. Мы представим особенно простое изложение такой несостоятельности. Причиной феномена забывания известных научных фактов является изменение требований к публикациям в научных журналах — ссылки только на современные публикации (не старше 5 лет).

Ключевые слова *модель Дирихле, байесовский вывод, состоятельность оценки, утеря знания, требования к цитированию*

DOI: 10.36535/0548-0027-2022-04-1

ВВЕДЕНИЕ

Весной 2019-2020 учебного года автор настоящей статьи согласился прочитать курс лекций «Байесовские методы в статистике и машинном обучении» для студентов – магистров отделения интеллектуальных систем Российского государственного гуманитарного университета (РГГУ). Этот курс был подготовлен на основе классического учебника [1] и включал в себя стандартный материал. Но для погружения студентов в современные исследования были проанализированы актуальные темы в байесовском машинном обучении, а именно: рассмотрено несколько отечественных курсов по байесовскому машинному обучению (Д.П. Ветрова (НИУ-ВШЭ/ШАД Яндекса), С.Н. Николенко (СПбГУ), Е.В. Бурнаева (СколТех)) и большое число курсов в Западных университетах.

Большинство из этих курсов несколько заключительных лекций посвящало байесовскому выводу для процессов Дирихле. Это материал обладает богатой математической структурой (китайский ресторан, переламывания палочки). Однако автор обнаружил факт неадекватности этого подхода – несостоятельность байесовского вывода, т. е. возможность смещения распределения в сторону от истинностного значения по мере накопления информации (увеличения объема обучающей выборки).

К сожалению, это обстоятельство совершенно не обсуждается в рассмотренных курсах, хотя статья [2], опубликованная на первых страницах тома 14 «Annals of Statistics» за 1986 г., содержит конкретные примеры такой несостоятельности именно для байесовского оценивания процессов Дирихле. Более того, на нескольких последующих страницах этого тома даются дополнительные комментарии известных ученых на эту же тему.

В настоящей статье обсудим феномен утери научного знания в частном случае и представим, на наш взгляд, правдоподобное объяснение такого явления.

НЕСОСТОЯТЕЛЬНОСТЬ БАЙЕСОВСКОГО ВЫВОДА

Чтобы убедить своих студентов в важности феномена несостоятельности, я предположил рассмотреть ситуацию в особенно простом случае, когда обучающие точки (на прямой) независимо порождаются нормальным законом распределения с единичной дисперсией и нулевым средним. Это тот случай, с которым так или иначе сталкивается каждый исследователь при статистическом анализе экспериментального материала.

Единственным дополнительным условием было то, что более общая модель допускает смеси нормальных законов с единичной дисперсией, причем компоненты задают мультиномиальное распределение, а внутри каждого компонента есть свое среднее, которое извлекается независимо из стандартного нормального распределения. Это и есть один из вариантов процессов Дирихле. Установленный мной результат показывает, что байесовский вывод будет приводить к заключению, что компонент не один, даже если обучающая выборка соответствует единственному компоненту! Содержательно это гласит, что количество компонентов смеси является гиперпараметром и не должно оцениваться по обучающей выборке. В противном случае, мы сталкиваемся с эффектом переподгонки модели (переобучения).

Несостоятельность байесовского вывода (для параметров сдвига) для процессов Дирихле была установлена и опубликована в основополагающей работе [2] в 1986 г., представившая дискуссию с несколькими ключевыми исследователями байесовской статистики. Более того, как я обнаружил уже после завершения лекций 2020 г., имеется работа [3], выложенная в общеизвестный архив препринтов в 2013 г., но неопубликованная ни в одном журнале и тоже оставшаяся незамеченной. В ней доказывается некоторое расширение установленного мной результата. Впрочем, мой вариант доказательства является более простым, так как он использует только явное вычисление средних и формулу полной вероятности для средних.

Рассмотрим составную модель распределения точек на прямой. Сначала для множества индексов $\{1, 2, \ldots, n\}$ обучающих примеров задается мультиномиальное распределение на его упорядоченных разбиениях

$$A = \langle A_1, A_2, \ldots, A_t\rangle: p(A)=(n_1\text{-}1)!\ldots(n_t\text{-}1)!/n!t!$$

(где $|A_j| = n_j$, $n = n_1 + \ldots + n_t$, $t!$ появляется в знаменателе из-за перестановок t кластеров разбиения; $n!$ – из-за перестановок всех элементов, а $(n_1\text{-}1)!\ldots(n_t\text{-}1)!$ в числителе – из-за перестановок внутри кластеров), этот коэффициент умножается на произведения плотностей элементов, чтобы образовать соответствующие степени (из-за независимости).

При заданном разбиении на кластеры зададим условное распределение центров кластеров по нормальному закону с единичной дисперсией и нулевым средним независимо друг от друга

$$p(\theta_1, \ldots, \theta_t \mid A) = \prod_{j=1}^{t} [\exp(-\theta_j^2/2)/(2\pi)^{1/2}].$$

Наконец, зададим условное распределение элементов каждого кластера независимо друг от друга и элементов других кластеров по нормальному закону с единичной дисперсией вокруг центра соответствующего кластера

$$p(x_1, \ldots, x_n \mid A, \theta_1, \ldots, \theta_t) = \prod_{j=1}^{t} \prod_{i} [\exp(-(x_i-\theta_j)^2/2)/(2\pi)^{1/2}]$$

(где $i \in A_j$).

Чтобы объяснить название «Байесовский вывод для процессов Дирихле», напомним, что распределение Дирихле является сопряженным в экспоненциальном классе с мультиномиальным распределением подобно тому, как бета-распределение сопряжено с биномиальным.

Обозначим через T случайное число кластеров. Тогда по теореме Байеса

$$p(T = 1| X_1, \ldots, X_n) = p(X_1, \ldots, X_n, T = 1)/$$
$$\Sigma_t\, p(X_1, \ldots, X_n, T = t) \le p(X_1, \ldots, X_n, T = 1)/$$
$$(p(X_1, \ldots, X_n, T = 1)+p(X_1, \ldots, X_n, T = 2)) = 1/$$
$$(1+p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1)).$$

Рассмотрим частный случай этой модели, в котором имеется только один кластер – это последовательность независимых нормально распределенных случайных величин (с нулевым средним и единичной дисперсией). Это классическая модель данных, для которой хотелось бы получить однозначные результаты (при возрастании обучающей выборки $n \to \infty$). К сожалению, байесовский вывод о числе кластеров несостоятелен.

Несостоятельность означает, что при возрастании объема n обучающей выборки (из единственного кластера) апостериорная вероятность числа кластеров не может стремиться к 1 с вероятностью 1.

Дробь в знаменателе

$$1/(1+p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1))$$

только уменьшится, если числитель ограничить случаем, когда в одном кластере только один обучающий пример, а остальные — в другом. Выделяя полные квадраты, с использованием нормального интеграла (для $S_{-k} = X_1+\ldots+X_n\text{-}X_k$) получаем

$$p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1) \ge (n+1)^{1/2}/$$
$$2(n\text{-}1)(2n)^{1/2}\, \Sigma_k \exp\,[S^2_{-k}/2(n+1) + (n^2+n\text{-}2)X^2_k/$$
$$4n(n+1) - S_{-k}\cdot X_k/n(n+1)].$$

Подробный вывод этого неравенства приводится также в работе [3]. Однако дальнейшие вычисления из этой статьи можно упростить, сведя к прямым вычислениям, что мы кратко наметим.

Среднее суммы равно сумме средних, причем средние всех слагаемых одинаковы. Так как обучающая выборка $X_1, \ldots, X_n$ – независимые стандартные нормальные случайные величины, то X_k и S_{-k} независимы, причем S_{-k} имеет нормальное распределение со средним 0 и дисперсией n-1. Из-за одинаковости средних можно рассмотреть одно из слагаемых

$$\exp [S^2_{-k}/2(n+1) + (n^2+n-2)X^2_k/4n(n+1) - S_{-k}\cdot X_k/n(n+1)]$$

(у которого отбросим нижние индексы при X и S).

Тогда

$$\mathbf{E}_S \mathbf{E}_X [\exp(S^2/2(n + 1) + (n^2 + n - 2)X^2/ 4n(n + 1) - S\cdot X/n(n + 1))] = \mathbf{E}_S [2^{½}n^{½}(n + 1)^{½}/ (n^2 + n+2)^{½}\cdot\exp \{(n^2 + 2)S^2/2n(n^2 + n + 2)\}] = = n\cdot(n + 1)^{1/2}/(n^2+1)^{1/2}.$$

Эти равенства получаются, выделяя полные квадраты, с использованием нормального интеграла, например, первое равенство

$$\mathrm{E}_X [\exp \{S^2/2(n + 1) + (n^2 + n - 2)X^2/ 4n(n + 1) - S\cdot X/n(n + 1))\}] = = \int dx\cdot\exp \{[-(x^2 - S^2/(n + 1) - (n^2 + n - 2)\, x^2/ (2n(n + 1)) + 2S\cdot x/(n(n + 1)]/2\}/(2\pi)^{½} = =\int dx\cdot\exp \{-(n^2 + n + 2)/ 4n(n + 1)\cdot(x - 2S/(n^2 + n + 2)^2\} + (n^2 + 2)S^2\}/ 2n(n^2 + n + 2)\}/ (2\pi)^{½} = (2)^{½}n^{½}\,(n + 1)^{½}/ (n^2 + n+2)^{½}\cdot\exp [(n^2 + 2)S^2/2n(n^2 + n + 2)$$

с применением равенства $\int dz\cdot\exp \{-z^2/2a\} = (2\pi a)^{½}$ при $a = 2n(n + 1)/(n^2 + n + 2)$.

Аналогичным образом вычисляется среднее E_S по S.

Усредняя по $X_1, \ldots, X_n$ и собирая вместе все сомножители, получаем

$$\mathrm{E}[p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1)] \geq \geq n\cdot(n+1)^{1/2}/2(n-1)(2n)^{1/2}\cdot n(n + 1)^{1/2}/(n^2+1)^{1/2},$$

где первый сомножитель n равен числу слагаемых.

Таким образом среднее дроби

$$p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1)$$

ограничено снизу $n^{1/2}/2^{3/2}$. Значит, существует множество ненулевой меры бесконечных последовательностей $X_1, \ldots, X_n, \ldots$, на которых предел дроби

$$p(X_1, \ldots, X_n, T = 2)/p(X_1, \ldots, X_n, T = 1)$$

бесконечно возрастает. Это и означает несостоятельность, так как на таких последовательностях апостериорная вероятность наличия одного кластера стремится к 0, тогда как все такие последовательности извлечены из единственного кластера (да еще с нулевым средним!).

ПРИЧИНЫ УТЕРИ НАУЧНОГО ЗНАНИЯ

Как же могло случиться так, что результат о несостоятельности байесовского вывода для процессов Дирихле, опубликованный в главном журнале по математической статистике, оказался полностью забытым?

Ответом, на мой взгляд, является порочная практика оформления пристатейных списков использованной литературы: редакционные коллегии рекомендуют (через рецензентов) включать в статьи ссылки на работы других авторов, опубликованные в том же журнале, причем нужно стараться использовать относительно свежие публикации, не старше 5 лет.

Такое поведение редколлегий очень рационально: цитирование свежих работ приводит к увеличению индексов цитируемости издания и, как следствие, к более высокому рейтингу в международных базах цитирования (*Scopus, Web of Science* и др.).

Ясно, что уже через 10 лет молодые исследователи будут лишены возможности приобрести знания, которые были общеизвестны: через 5 лет исчезнут ссылки, а еще через 5 лет наставники не смогут передать это и в устной форме. Сами же начинающие исследователи задавлены необходимостью отслеживать новые результаты в своей области, скорость публикаций которых сейчас превысила, на мой взгляд, возможности восприятия молодого исследователя, если ему не посчастливилось оказаться под опекой квалифицированного старшего коллеги.

ЗАКЛЮЧЕНИЕ

Цель настоящей статьи – обратить внимание общественности и руководства наукой на видимые невооруженным взглядом изъяны в современной практике оценивания научных журналов по рейтингу цитирования и попаданию в высокие квартили международных баз.

Следует отметить, что, будучи членом редакционной коллегии нескольких журналов, я рассматриваю источник проблемы с точки зрения издателя. При этом точка зрения автора является его личным мнением и может не совпадать с мнениями членов редакционных коллегий журналов, в которых он состоит.

Для исследователя ситуация выглядит более благоприятной благодаря наличию базы препринтов arXiv, но тут возникает вопрос доверия к опубликованным там статьям, так как они не проходят такого серьезного научного рецензирования, как в классических изданиях.

* * *

Автор благодарит проф. Р.С. Гиляревского (ВИНИТИ РАН) за поддержку и полезные обсуждения и проф. М.С. Гельфанда (ИППИ РАН) за полезные дискуссии.

СПИСОК ЛИТЕРАТУРЫ

1. Bishop C.M. Pattern Recognition and Machine Learning. – NY: Springer, 2006. – 738 p.
2. Diaconis P., Freedman D.A. On the consistency of Bayes estimates (with discussion) // Annals of Statistics. – 1986. – Vol. 14. – P. 1-26.
3. Miller J.W., Harrison M.T. A simple example of Dirichlet process mixture inconsistency for the number of components // *arXiv*:1301.2708v1. – 2013. – P. 1-8.

Материал поступил в редакцию 16.02.22

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – доктор физико-математических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН
e-mail: vinogradov.d.w@gmail.com