Федеральное государственное автономное образовательное учреждение высшего образования
“Московский физико-технический институт (национальный исследовательский университет)” (МФТИ)

На правах рукописи
УДК 519.22

Волков Никита Алексеевич

О некоторых свойствах вероятностных распределений и их применении в задачах машинного обучения

05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Москва — 2020

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования "Московский физико-технический институт (национальный исследовательский университет)", на кафедре анализа данных физтех-школы прикладной математики и информатики

Научный руководитель:

Жуковский Максим Евгеньевич,
доктор физико-математических наук,
доцент кафедры дискретной математики
физтех-школы прикладной математики и информатики;
Московского физико-технического института;

Ведущая организация:

Федеральное государственное бюджетное учреждение науки
Федеральный исследовательский центр
"Коми научный центр Уральского отделения
Российской академии наук"

Защита диссертации состоится «15» декабря 2020г. в 12:00 на заседании диссертационного совета ФПМИ.05.13.17.002 по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9.

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национального исследовательского университета):
`https://mipt.ru/education/post-graduate/soiskateli-fiziko-matematicheskie-nauki.php`.

Работа представлена «1» октября 2020 г. в Аттестационную комиссию Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание учёной степени кандидата наук, доктора наук в соответствии с п. 3.1 ст. 4 Федерального закона «О науке и государственной научно-технической политике».

Актуальность и степень разработанности темы.

1. Для биномиальной случайной величины ξ с параметрами $n \in \mathbb{N}$ и b/n хорошо известно, что ее медиана равна b, если $b \in \{1, \ldots, n\}$[1]. Рассмотрим биномиальную случайную величину $\xi_{b,n,c}$ с параметрами n и $\frac{b}{n+c}$, где $b < n$ — натуральные числа и $c \in (0, 1)$. Обозначим $p_{b,n,c} := \mathsf{P}(\xi_{b,n,c} < b)$. Ранее были доказаны[2] следующие утверждения:

- если $n \geqslant 3b + 2$, то $p_{b+1,n,0} > p_{b,n,0}$;
- если $n \leqslant 3b + 1$, то $p_{b+1,n,0} < p_{b,n,0}$.

Заметим, что в 1968 году Джогдео и Самуэльс[3] изучали поведение вероятности $p_{b,n,0}$, а также отношения $\mathsf{P}(\xi = b)$ и $1/2 - \mathsf{P}(\xi < b)$. Такого рода вопросы мотивированы известным вопросом Рамануджана, относящимся к пуассоновским случайным величинам[4].

Кроме того, исследование монотонности $p_{b,n,c}$ по b мотивировано задачей о неравенстве малых отклонений[5], которая может быть сформулирована следующим образом: для $c > 0$ найти минимум $\mathsf{P}(\xi_1+\xi_2+...+\xi_n < n+c)$ по всем множествам независимых неотрицательных случайных величин $\{\xi_1, ..., \xi_n\}$ с одинаковым средним. Эта задача до сих пор не решена. Тем не менее, было показано[6], что оптимальными случайными величинами являются величины, принимающие два значения с вероятностью 1 (как говорится, *с двумя атомами*). Если мы далее ограничимся *одинаково распределенными* случайными величинами с двумя атомами, то сведем исходную задачу к анализу монотонности $p_{b,n,c}$ по b.

В настоящей диссертации исследована монотонность $p_{b,n,c}$ по b для произвольного $c \in [0, 1]$.

2. Свойства распределения Стьюдента впервые исследовал Уильям Госсет. Он обратил внимание, что стандартизированное (отцентрированное и отмасштабированное) выборочное среднее нормальной выборки при замене

[1]Lord N.. Binomial averages when the mean is an integer // The Mathematical Gazette. — 2018. — V. 94. — P. 331–332.

[2]Dmitriev D., Zhukovskii M.. On monotonicity of Ramanujan function for binomial random variables — 2018. — arXiv:1807.06527

[3]Choi, K.P. (1994). On the medians of Gamma distributions and an equation of Ramanujan. // Proc. Amer. Math. Soc. — 1994. — V. 121. — P. 245–251.

[4]Jogdeo, K., Samuels, S.M.. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation // Ann. Math. Statist. — 1968. — V. 39. — P. 1191–1195.

[5]Feige, U.. On Sums of Independent Random Variables with Unbounded Variance and Estimating the Average Degree in a Graph. // SIAM J. Comput.. — 2006. — V. 35. — P. 964–984.

[6]He, S., et al.. Bounding Probability of Small Deviation: A Fourth Moment Approach // JSTOR. — 2010. — V. 35. — P. 208–232.

неизвестной дисперсии на ее оценку имеет распределение, отличное от нормального. В литературе[7] приводятся некоторые свойства распределения Стьюдента. Например, вычисляется плотность распределения с помощью определения случайной величины с таким распределением через комбинацию случайных величин, имеющих нормальное распределение и гамма-распределение, а также его математическое ожидание и дисперсия. Естественным образом, подобно нормальному распределению, вводится многомерный аналог распределения Стьюдента. Кроме того, известно[8], что условное распределение компонент вектора, имеющего многомерное распределение Стьюдента, также имеет распределение Стьюдента.

Для описания данных часто используют смеси нормальных распределений. Оценка параметров такой смеси происходит с помощью EM-алгоритма. При наличии выбросов в данных логично рассматривать смесь распределений Стьюдента. Идеи оценки параметров смеси распределений Стьюдента были описаны ранее[9]. Оценку параметров смеси распределений Стьюдента можно также производить с помощью методов MCMC[10].

Наряду с выбросами другой достаточно распространенной проблемой анализа данных является наличие в них пропущенных значений. Известны[11] различные методы работы с пропущенными значениями, в частности, итерационные процедуры оценки параметров смеси нормальных распределений при наличии пропусков, а также для логлинейных моделей. Также ранее рассмотрен[12] метод построения генеративных топографических карт (GTM) при наличии пропущенных данных, основанный на EM-алгоритме. В качестве статистической модели рассматриваются смеси регрессионных моделей, использующих шарообразные распределения Стьюдента. Иными словами, матрица ковариаций этих распределений имеет диагональный вид с одинаковыми значениями на диагонали. Этот случай можно считать частным случаем рассматриваемой в диссертации задачи.

В настоящей диссертации разработана итерационная процедура оценки параметров смеси многомерных распределений Стьюдента по выборкам, в которых могут присутствовать пропущенные значения.

---

[7]Козлов М.В., Прохоров А.В. Введение в математическую статистику. Москва : МГУ, 1987.

[8]Kibria B.M.G., Joarder A.H. A short review of multivariate t-distribution // Journal of Statistical Research. 2006. V. 40. P. 256–422.

[9]Peel D., Mclachlan G. Robust Mixture Modelling Using the t-distribution // Statistics and Computing. 2000. V. 10. P. 339–348.

[10]Fruhwirth-Schnatter S. Finite Mixture and Markov Switching Models. New York : Springer, 2006.

[11]Roderick J Little, Donald B. Rubin Statistical Analysis with Missing Data. Hoboken, New Jersey : y John Wiley & Sons, Inc., 2002.

[12]A. Vellido Missing data imputation through GTM as a mixture of t-distributions // Neural Networks. 2006. V. 19. P. 1624–1635.

3. Существует ряд методов оценки представительности проб пластовых флюидов, таких как

- проверка герметичности пробоотборных камер;
- сопоставление давления насыщения нефти с давлением сепарации при температуре сепарации и др.;
- метод Хоффмана-Крампа-Хоккота, основанный на корреляции констант равновесия;
- определение представительности проб по критерию загрязненности технологическими жидкостями, применяемыми при бурении, перфорации и освоении скважины.

В условиях же, когда имеются только сырые данные, выше представленные методы не могут быть применены, в связи с чем возникает необходимость в разработке алгоритмов выявления потенциально некорректных значений по сырым данным.

Задача предсказания PVT-свойств методами машинного обучения ранее рассматривалась в сильно ограниченном варианте. Например, рассматривалось[13] предсказание давления насыщения через другие свойства с помощью нейронных сетей. Аналогичным образом рассматривались[14] предсказания объемного коэффициента нефти. В статье[15] для предсказания упомянутых выше признаков используется SVM-регрессия.

В настоящей диссертации предложен метод машинного обучения на основе рассмотренной в диссертации оценки параметров смеси распределений Стьюдента, а также реализующий его программный продукт, позволяющий значительно расширить спектр решаемых задач и повысить точность решений рассматриваемых ранее задач.

Цель работы и задачи исследования.

1. Исследование монотонности величины $p_{b,n,c}$ по b для произвольного числа $c \in (0, 1]$.

[13] Alakbari F., Elkatatny S., Baarimah S. Prediction of Bubble Point Pressure Using Artificial Intelligence AI Techniques // Proc. of the SPE Middle East Artificial Lift Conference and Exhibition. 2016. 10.2118/184208-MS.

[14] Osman E.A., Abdel-Wahhab O.A., Al-Marhoun M.A. Prediction of Oil PVT Properties Using Neural Networks // Society of Petroleum Engineers. 2001. doi:10.2118/68233-MS.

[15] El-Sebakhy E.A., Sheltami T., Al-Bokhitan S.Y., Shaaban Y., Raharja P.D., Khaeruzzaman Y. Support Vector Machines Framework for Predicting the PVT Properties of Crude Oil Systems // Society of Petroleum Engineers. 2007. doi:10.2118/105698-MS.

2. Разработка процедуры оценки параметров смеси многомерных распределений Стьюдента по выборке, в которой присутствуют пропущенные значения.

3. Разработка инструментов для комплексной оценки достоверности данных исследований PVT-свойств пластовых флюидов.

Структура диссертации.

Диссертация состоит из введения, 3 глав, заключения и библиографии. Общий объем диссертации 102 страниц, из них 95 страниц текста (не считая титульного листа, оглавления и библиографии), на которых приведены 9 рисунков и 4 таблицы. Библиография включает 37 наименований на 4 страницах.

В первой главе приводится обобщение результата Дмитриева и Жуковского для случайной величины ξ с параметрами $n \in \mathbb{N}$ и $b/(n+c)$ для произвольного $c \in [0, 1]$, а также подтверждена гипотеза, сформулированная ими. С помощью полученного инструмента можно сформулировать следствия, аналогичные этой гипотезе.

Во второй главе для выборки из многомерной смеси распределений Стьюдента предложен новый способ оценки параметров смеси на основе EM-алгоритма, в котором на E-шаге применяется вариационный байесовский вывод. Основной особенностью данного способа является то, что он позволяет получать оценки в случае наличия пропусков в данных.

В третьей главе на основе смеси распределений Стьюдента построен метод машинного обучения, позволяющий с помощью одной модели решать задачи регрессии по любому набору признаков, кластеризации, обнаружения аномалий. Каждая из этих задач может быть решена моделью при наличии пропусков в данных. На основе данного метода разработан инструмент для комплексной оценки достоверности данных исследований PVT-свойств пластовых флюидов. Приведены результаты тестирования на данных PVT-свойств, показано, что предсказания во многих случаях точнее широко известных методов машинного обучения по метрикам MAPE и RMSPE.

Научная новизна. Теоретическая и практическая значимость работы.

1. Получено обобщение результата Дмитриева и Жуковского, подтверждена сформулированная ими гипотеза. Данный результат носит теоретический характер. С помощью полученного инструмента можно сформулировать следствия, аналогичные этой гипотезе.

2. Впервые разработана итерационная процедура получения оценки параметров смеси многомерных распределений Стьюдента по выборкам, в которых имеются пропущенные значения. Данный результат носит как теоретический, так и практический характер. Полученный метод можно обобщать на более сложные вероятностные модели. Также данный метод применен для решения задач машинного обучения.

3. Предложен метод машинного обучения на основе смеси многомерных распределений Стьюдента, позволяющий решать задачи кластеризация, регрессии, детектирования аномалий, в том числе, при наличии пропусков в данных. С помощью данного метода получены принципиально новые инструменты для оценки достоверности данных исследований PVT-свойств пластовых флюидов. Данный результат носит практический характер. Инструмент используется в работе геологов.

Положения, выносимые на защиту.

1. Если $c = 1$, то $p_{b+1,n,c} > p_{b,n,c}$ при любых $1 \leqslant b < n$.

2. Если $n \geqslant 3b + 2$, то $p_{b+1,n,c} > p_{b,n,c}$.

3. Порог, при котором монотонность меняется, равен $\frac{n}{3(1-c)}(1 + o(1))$. Формально, $\forall \varepsilon > 0\ \forall \delta > 0\ \forall c \in (0,1)\ \exists n_0\ \forall n \geqslant n_0\ \forall b \in (\varepsilon n, n)$:

   - при $b < \frac{n(1-\delta)}{3(1-c)}$ выполнено $p_{b+1,n,c} > p_{b,n,c}$,
   - при $b > \frac{n(1+\delta)}{3(1-c)}$ выполнено $p_{b+1,n,c} < p_{b,n,c}$.

4. Разработанная в диссертации итерационная процедура получения оценки параметров в смеси многомерных распределений Стьюдента общего вида (без дополнительных ограничений на параметры) при помощи вариации EM-алгоритма, в которой на E-шаге применяется вариационный байесовский вывод, позволяет вычислять оценку параметров по выборке, в которых часть значений ненаблюдаема.

5. Разработанный в диссертации метод машинного обучения, основанный на смеси многомерных распределений Стьюдента, позволяет решать следующие задачи:

   (a) кластеризация точек на выбранное количество кластеров,

   (b) выявление аномальных точек,

   (c) регрессия для предсказания любого набора вещественных признаков при использовании любого другого набора (предсказание и доверительный интервал).

   Каждая из этих задач может быть решена моделью при наличии пропусков в данных.

6. Разработанный программный продукт, реализующий вышеупомянутый метод, является новым инструментом для оценки достоверности данных исследований PVT-свойств пластовых флюидов.

Методы исследования.

Для доказательства результатов первой главы диссертации широко применялся аппарат следующих дисциплин: теория вероятностей и математический анализ. Некоторые вычисления производились при помощи символьных вычислений Matlab. Для получения результатов второй главы применялся аппарат теории вероятностей, байесовских методов, линейной алгебры. Программные инструменты третьей главы разработаны на языке Python.

Степень достоверности и апробация результатов.

Основные результаты диссертации содержатся в 5 работах, приведенных в настоящего реферата. Три работы опубликованы в журналах, индексируемых Scopus, еще 2 работы — RSCI. Первые 4 работы написаны автором самостоятельно, остальные авторы принимали сопроводительное участие (постановка задач, экспертные советы, перевод на английский). В пятой работе автором написан раздел "Анализ PVT-свойств", который включен в главу 3 настоящей диссертации.

Результаты диссертации докладывались на следующих конференциях и семинарах:

1. Российская нефтегазовая техническая конференция SPE (2019),
2. 62-я научная конференция МФТИ (2019),
3. Научный семинар кафедры анализа данных МФТИ (2020),
4. Семинар отдела трудноизвлекаемых углеводородов МФТИ (2020).

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении излагается история исследований, относящихся к вероятностным распределениям, исследованию применения методов анализа данных к PVT-свойствам, а также описывается структура диссертации.

В главе 1 представлены полученные результаты и их доказательства по исследованию биномиального распределения. Для формулировки результатов введем биномиальную случайную величину $\xi_{b,n,c}$ с параметрами n и $\frac{b}{n+c}$, где $b < n$ — натуральные числа и $c \in (0,1)$. Обозначим $p_{b,n,c} := \mathsf{P}(\xi_{b,n,c} < b)$. Дмитриев и Жуковский доказали теорему о том, что если $n \geqslant 3b+2$, то $p_{b+1,n,0} > p_{b,n,0}$, а если $n \leqslant 3b+1$, то $p_{b+1,n,0} < p_{b,n,0}$.

В диссертации доказана следующая теорема.

ТЕОРЕМА 1. *Если* $c = 1$, *то* $p_{b+1,n,c} > p_{b,n,c}$ *при любых* $1 \leqslant b < n$.

Из монотонности $p_{b,n,1}$, получаемой из данной теоремы, следует приведенная далее гипотеза, ранее сформулированная Дмитриевым и Жуковским.

ГИПОТЕЗА 1. *Пусть* $\alpha \in [0,1), \beta > 1$. *Пусть* b — *целое число такое, что*

$$b < \frac{n+1-n\alpha}{\beta-\alpha} \leqslant b+1.$$

Тогда $\mathsf{P}(\xi_1 + ... + \xi_n < n+1) \geqslant p_{b,n,1}$, *где* $\xi_1, ..., \xi_n$ — *независимые одинаково распределенные двухатомные случайные величины со значениями* α *и* β, *средним 1, и равенство выполняется тогда и только тогда, когда* $\alpha = 0$ *и* $\frac{n+1}{\beta} = b+1$.

Аналогичные этой гипотезе утверждения можно сформулировать для любого $c \in (0,1)$. В этой связи в данной статье исследована монотонность для всех $c \in (0,1)$.

Во-первых, обобщен первый пункт результата Дмитриева и Жуковского на случай произвольного $c \in [0,1]$.

ТЕОРЕМА 2. *Если* $n \geqslant 3b+2$, *то* $p_{b+1,n,c} > p_{b,n,c}$.

Кроме того, получен *асимптотический* результат, утверждающий, что порог, при котором монотонность меняется, равен $\frac{n}{3(1-c)}(1+o(1))$.

ТЕОРЕМА 3. $\forall \varepsilon > 0\ \forall \delta > 0\ \forall c \in (0,1)\ \exists n_0\ \forall n \geqslant n_0\ \forall b \in (\varepsilon n, n)$:

- *при* $b < \frac{n(1-\delta)}{3(1-c)}$ *выполнено* $p_{b+1,n,c} > p_{b,n,c}$

- *при* $b > \frac{n(1+\delta)}{3(1-c)}$ *выполнено* $p_{b+1,n,c} < p_{b,n,c}$

Благодаря этому существенно обобщен результат Дмитриева и Жуковского и получен инструмент, с помощью которого можно сформулировать следствия, аналогичные гипотезе 1.

Доказательства всех приведенных теорем основываются на нескольких утверждениях, для формулировки которых введем следующие обозначения:

1. $\Delta_{b,n+c} = \left[1 - \frac{b+1}{n+c}, 1 - \frac{b}{n+c}\right]$;
2. $g(z) = (1-z)^{b-1} z^{n-b}$;
3. $g_{b,n,c} = \int_{\Delta_{b,n+c}} g(z)dz$.

Утверждение 1 из статьи Дмитриева и Жуковского обобщается на случай произвольного $c \in [0,1]$ в следующем виде.

УТВЕРЖДЕНИЕ 1.

$$p_{b+1,n,c} - p_{b,n,c} = \frac{\left(\frac{b+1}{n+c}\right)^b \left(1 - \frac{b+1}{n+c}\right)^{n-b} - b\int_{1-\frac{b+1}{n+c}}^{1-\frac{b}{n+c}} (1-z)^{b-1} z^{n-b} dz}{b\int_0^1 (1-z)^{b-1} z^{n-b} dz}.$$

Далее используя утверждение 2 из статьи Дмитриева и Жуковского и формулу Тейлора с остаточным членом в форме Лагранжа, в диссертации выводится верхняя и нижняя оценки для $g(z)$ на $\Delta_{b,n+c}$. Для формулировки данного утверждения обозначим для любого $\ell \in \mathbb{Z}_+$

$$g_\ell(z) = \frac{1}{\ell!}\left(z - 1 + \frac{b+1}{n+c}\right)^\ell \left[\frac{\partial^\ell g}{\partial z^\ell}\left(1 - \frac{b+1}{n+c}\right)\right]$$

ℓ-й член разложения g по формуле Тейлора, и $g_{\leqslant 3}(z) = \sum_{\ell=0}^{3} g_\ell(z)$.

УТВЕРЖДЕНИЕ 2. *Для* $5 \leqslant b \leqslant n/3$ *и всех* $z \in \Delta_{b,n+c}$,

$$g_{\leqslant 3}(z) + g_4^+(z) \geqslant g(z) \geqslant g_{\leqslant 3}(z) + g_4^-(z),$$

где

$$g_4^-(z) = \frac{1}{24}\left(z - 1 + \frac{b+1}{n+c}\right)^4 d_4^-(z), \quad g_4^+(z) = \frac{1}{24}\left(z - 1 + \frac{b+1}{n+c}\right)^4 d_4^+(z),$$

$$d_4^-(z) = \frac{\partial^4 g}{\partial z^4}\left(1 - \frac{b+1}{n+c}\right), \quad d_4^+(z) = \frac{\partial^4 g}{\partial z^4}\left(1 - \frac{b}{n+c}\right).$$

Из доказательства утверждения 2, получаем, что для любого $5 \leqslant b \leqslant n/2$, функции d_4^- и d_4^+ являются нижними и верхними оценками $\partial^4 g/\partial z^4$ на $\Delta_{b,n+c}$.

Приведем кратко основные шаги доказательств теорем, сформулированных в главе 1.

Доказательство теоремы 2 при помощи утверждения 2 сводится к доказательству положительности некоторого многочлена $F(n,b,c)$, определение которого дано в диссертации. Его положительность показывается путем исследования монотонности первых и вторых производных по c и их значений в крайних точках. В результате установлено, что $F(n,b,c)$ возрастает по c, откуда следует утверждение теоремы, поскольку положительность $F(n,b,0)$ доказана Дмитриевым и Жуковским.

Доказательство теоремы 1 при помощи утверждения 2 сводится к доказательству положительности некоторого многочлена $P(n,b,c)$, определение которого дано в диссертации. Далее с помощью замены $n = (b+1)x$, учитывая, что $n/(b+1) \geqslant 1$, исследуется на неотрицательность многочлен $F(b,x,d) = P((b+1)x,b,d)$ от независимых переменных $b \in \mathbb{N}, x \geqslant 1, d \in [0,1]$.

Многочлен $F(b,x,d)$ является многочленом 9-й степени от b. Обозначим $F_k(x,d)$ — коэффициент в многочлене $F(b,x,d)$ перед b^k, тем самым получим $F(b,x,d) = \sum_{k=0}^{9} F_k(x,d)b^k$. В диссертации доказана неотрицательность при $b \geqslant 4$ коэффициентов $F_9(x,d)$, $F_8(x,d)$, $F_7(x,d)$, $F_6(x,d)$, $F_5(x,d)$, $F_2(x,d)$, $F_0(x,d)$, а также неотрицательность выражений $F_6(x,d)b^2+F_4(x,d)$, $F_5(x,d)b^2+F_3(x,d)$, $F_2(x,d)b+F_1(x,d)$. Случаи $F(1,x,d)$, $F(2,x,d)$, $F(3,x,d)$ исследуются отдельно.

При доказательстве теоремы 3 применена формула Тейлора с остаточным членом в форме Лагранжа, согласно которой для любого $z \in \left[1-\frac{b+1}{n+c}, 1-\frac{b}{n+c}\right]$ существует $d \in [0,1]$, для которого $g(z) = \sum_{\ell=0}^{3} g_\ell(z) + r(z,d)$, где

$$r(z,d) = \frac{1}{24}\left(z-1+\frac{b+1}{n+c}\right)^4 \frac{\partial^4 g}{\partial z^4}\left(1-\frac{b+d}{n+c}\right).$$

Далее по утверждению 1 доказательство сводится к исследованию знака многочлена $P(n,b,c,d)$. Этот многочлен при $b \in (\varepsilon n, n)$ и $n \to +\infty$ асимптотически эквивалентен $b^4n^4H(b/n)$, где $H(x) = 20 + (60c-120)x + (-180c+240)x^2 + (180c-200)x^3 + (-60c+60)x^4$. Иначе говоря, $\frac{P(n,b,c,d)}{b^4n^4H(b/n)} = 1 + o(1)$ при $n \to +\infty$. Многочлен $H(x)$ имеет корень $x = 1$ кратности 3 и корень $x = \frac{1}{3(1-c)}$. Тем самым асимптотически смена знака происходит при $b = \frac{n}{3(1-c)}$.

В **главе 2** для выборки из многомерной смеси распределений Стьюдента предложен новый способ оценки параметров смеси на основе EM-алгоритма, в котором на E-шаге применяется вариационный байесовский вывод. Основной особенностью данного способа является то, что он позволяет получать оценки в случае наличия пропусков в данных.

Плотность в модели смеси многомерных распределений Стьюдента имеет вид

$$p(x) = \sum_{j=1}^{k} w_j p(x|\mu_j, \Sigma_j, \nu),$$

где $p(x|\mu_j, \Sigma_j, \nu)$ — плотность многомерного распределения Стьюдента с ν степенями свободы с центром в точке μ_j и матрицей масштаба Σ_j. Параметр ν является гиперпараметром модели, то есть алгоритм не подбирает его значение.

Для вывода параметров используется следующее представление распределения Стьюдента. Пусть случайный вектор $\xi \sim \mathcal{N}(0, \Sigma)$ и случайная величина $\eta \sim \Gamma(\nu/2, \nu/2)$ независимы, а $\mu \in \mathbb{R}^d$ — некоторый фиксированный вектор. Тогда случайный вектор $X = \mu + \xi/\sqrt{\eta}$ имеет распределение $T_\nu(\mu, \Sigma)$.

Пусть $X = (X_1, ..., X_n)$ — выборка векторов из такой смеси распределений, причем некоторые значения в этой выборке могут быть неизвестными. Предлагаемый способ оценки параметров смеси заключается в выборе некоторого начального приближения на параметры (и вычисления начального приближения зависящих от них величин) и выполнения следующих шагов на каждой итерации.

1. Для каждого объекта X_i введем номер кластера в виде $T_i = (T_{i1}, ..., T_{ik}) \in \{0, 1\}^k$, причем $\sum_{j=1}^{k} T_{ij} = 1$. Величина $T_{ij} = 1$ если объект X_i взят из кластера j и $T_{ij} = 0$ иначе. Обозначим $T = (T_1, ..., T_n)$.

2. Также для каждого объекта X_i введем случайную величину Y_i, имеющую распределение $\Gamma(\nu/2, \nu/2)$ таким образом, что справедливо представление

$$X_i = \sum_{j=1}^{k} \left(\mu_j + \xi_{ij} \Big/ \sqrt{Y_i} \right) I\{T_{ij} = 1\},$$

где случайный вектор ξ_{ij} имеет распределение $\mathcal{N}(0, \Sigma_j)$. Обозначим $Y = (Y_1, ..., Y_n)$.

Поскольку некоторые компоненты некоторых векторов X_i могут быть неизвестными, разделим также выборку X на наблюдаемые и скрытые величины. Пусть $J_i \subset \{1, ..., d\}$ — индексы известных компонент вектора X_i.

Тогда обозначим $X_{i,(k)} = (X_{ij})_{j \in J_i}$ и $X_{i,(u)} = (X_{ij})_{j \notin J_i}$ — векторы известных и неизвестных значений компонент вектора X_i соответственно. Обозначим также $X_{(k)} = (X_{1,(k)}, ..., X_{n,(k)})$ и $X_{(u)} = (X_{1,(u)}, ..., X_{n,(u)})$ — совокупности всех известных и неизвестных значений в выборке соответственно, а $d_{i,(k)}$ и $d_{i,(u)}$ — размерности векторов $X_{i,(k)}$ и $X_{i,(u)}$ соответственно.

Далее будем использовать обозначения $w = (w_1, ..., w_k), \mu = (\mu_1, ..., \mu_k), \Sigma = (\Sigma_1, ..., \Sigma_k)$. EM-алгоритм с вариационным байесовским выводом данной задаче представляется в следующем виде.

E-шаг. Найти распределение $X_{(u)}$, T и Y при условии $X_{(k)}$ при текущих значениях w, μ, Σ. Это производится с помощью вариационного вывода, который приближает это распределение в классе распределений, при которых $X_{(u)}$, T и Y условно независимы, с помощью выполнения итераций

1. $\ln r(t) \propto \mathsf{E}_\gamma \mathsf{E}_u \ln p(X, t, Y | w, \mu, \Sigma, \nu)$.
2. $\ln p_u(x_{(u)}) \propto \mathsf{E}_r \mathsf{E}_\gamma \ln p(X_{(k)}, x_{(u)}, T, Y | w, \mu, \Sigma, \nu)$.
3. $\ln \gamma(y) \propto \mathsf{E}_r \mathsf{E}_u \ln p(X, T, y | w, \mu, \Sigma, \nu)$

M-шаг. Обновить значения w, μ, Σ.

$$\mathsf{E}_r \mathsf{E}_u \mathsf{E}_\gamma \ln p(X, T, Y | w, \mu, \Sigma, \nu) \to \max_{w, \mu, \Sigma}$$

В начале производится выбор некоторого случайного начального приближения для значений w, μ, Σ и на распределения $X_{(u)}$, T и Y. По текущим приближениям апостериорных распределений T, Y и $X_{(u)}$ на каждом шаге используются следующие величины и функции

1. $r_{ij} = \mathsf{P}_r(T_{ij} = 1)$
2. $\varphi_i(\mu, \Sigma) = \mathsf{E}_u (X_i - \mu)^T \Sigma^{-1} (X_i - \mu)$
3. $c_i = \mathsf{E}_\gamma Y_i, \gamma_i = \mathsf{E}_\gamma \ln Y_i$,

Приведем общие формулы получаемой процедуры после упрощения.

E-шаг. Выполнить несколько итераций следующих трех шагов:

I. Вычислить некоторые вспомогательные величины

$$r_{ij} = \frac{w_j \exp\left[-\frac{1}{2} c_i \, \varphi_i(\mu_j, \Sigma_j)\right] / \sqrt{\det \Sigma_j}}{\sum\limits_{s=1}^{k} w_s \exp\left[-\frac{1}{2} c_i \, \varphi_i(\mu_s, \Sigma_s)\right] / \sqrt{\det \Sigma_s}},$$

Аналогично смеси нормальных распределений, величина r_{ij} отвечает за вероятность того, что объект X_i был получен из j-й компоненты смеси при текущем приближении параметров w_j, μ_j, Σ_j.

II. Для каждого объекта i, без ограничения общности считая, что вектор X_i и параметры его распределения можно представить в виде

$$X_i = \begin{pmatrix} X_{i,(k)} \\ X_{i,(u)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_{i,(k)} \\ \mu_{i,(u)} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Lambda_{i,(kk)} & \Lambda_{i,(ku)} \\ \Lambda^T_{i,(ku)} & \Lambda_{i,(uu)} \end{pmatrix},$$

разделив их тем самым на известные и неизвестные значения соответственно, вычислить

$$\alpha_{ij} = -\frac{r_{ij} c_i}{2}, \qquad B_i = \sum_{j=1}^{k} \alpha_{ij} \Lambda_{j,(uu)},$$

$$a_i = \sum_{j=1}^{k} \alpha_{ij} \left(\Lambda^T_{j,(ku)} (X_{i,(k)} - \mu_{j,(k)}) - \Lambda_{j,(uu)} \mu_{j,(u)} \right).$$

Переопределить функции φ_i как

$$\varphi_i(\mu, \Sigma) = (\widehat{X}_i - \mu)^T \Sigma^{-1} (\widehat{X}_i - \mu) + \operatorname{tr}\left[B_i^{-1} \Lambda_{(uu)} \right],$$

где

$$\widehat{X}_i = \begin{pmatrix} X_{i,(k)} \\ B_i^{-1} a_i \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_{(k)} \\ \mu_{(u)} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Lambda_{(kk)} & \Lambda_{(ku)} \\ \Lambda^T_{(ku)} & \Lambda_{(uu)} \end{pmatrix}.$$

Вектор $B_i^{-1} a_i$ и матрица B_i^{-1} определяют параметры нормального распределения величины $X_{i,(u)}$.

III. Вычислить

$$v_i = \frac{\nu + d}{2}, \quad b_i = \frac{\nu}{2} + \frac{1}{2} \sum_{j=1}^{k} r_{ij} \ (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j),$$

$$c_i = b_i / v_i, \quad \gamma_i = \psi(b_i) - \ln v_i$$

где d — размерность пространства признаков. Величины v_i и b_i имеют смысл параметров апостериорного гамма-распределения на гамма-распределенные величины из определения распределения Стьюдента при текущем приближении остальных параметров, а c_i и γ_i — соответствующие математические ожидания этих величин и их логарифмов.

М-шаг. Вычислить новое приближение параметров

$$w_j = \sum_{i=1}^{n} r_{ij} \Bigg/ \sum_{i,s=1}^{n,k} r_{is}, \quad \mu_j = \sum_{i=1}^{n} r_{ij} c_i \, \widehat{X}_i \Bigg/ \sum_{i=1}^{n} r_{ij} c_i,$$

$$\Sigma_j = \sum_{i=1}^{n} r_{ij} c_i \left[(\widehat{X}_i - \mu_j)(\widehat{X}_i - \mu_j)^T + \widehat{S}_i \right] \Bigg/ \sum_{i=1}^{n} r_{ij},$$

где

$$\widehat{S}_i = \begin{pmatrix} \mathbf{0}_{|J_i| \times |J_i|} & \mathbf{0}_{|J_i| \times (d - |J_i|)} \\ \mathbf{0}^T_{|J_i| \times (d-|J_i|)} & B_i^{-1} \end{pmatrix},$$

а $\mathbf{0}_{n \times m}$ — нулевая матрица размера $n \times m$.

В **главе 3** на основе смеси распределений Стьюдента приводится построение метода машинного обучения, который позволяет с помощью одной модели решать задачи регрессии по любому набору признаков, кластеризации, обнаружения аномалий. Каждая из этих задач может быть решена моделью при наличии пропусков в данных.

Рассмотрим каждую из них подробнее, предполагая модель смеси распределений, в которой плотность имеет вид

$$p(x) = \sum_{j=1}^{k} w_j p(x|\theta_j),$$

где θ_j — параметры распределения j-й компоненты смеси (например, вектор средних и матрица ковариаций).

Компоненты смеси можно рассматривать в качестве перекрывающихся кластеров. Каждый объект $x \in \mathbb{R}^d$ с некоторой вероятностью может быть отнесен к одному из кластеров. Условная вероятность того, что объект x соответствует кластеру j равна

$$p_j(x) = \frac{w_j p(x|\theta_j)}{\sum\limits_{s=1}^{k} w_s p(x|\theta_s)}.$$

Объект $x \in \mathbb{R}^d$ считается аномальным, если значение плотности $p(x)$ меньше некоторого порогового значения q. Величина q выбирается как значение плотности $p(x)$, при котором вероятность получить объект с плотностью, не превосходящей q, в точности равна 0.05. Иначе говоря, величина q является решением уравнения

$$\int\limits_{\mathbb{R}^d} p(x) I\{p(x) \leqslant q\} dx = 0.05.$$

Оба рассмотренных выше метода работают только в случае, если в объекте $x \in \mathbb{R}^d$ известны все значения, то есть отсутствуют пропуски. При наличии пропусков плотность объекта x можно оценить как интеграл плотности по подпространству пропущенных значений. Подробнее, пусть $x_{(k)}$ — вектор известных значений объекта, а $x_{(u)}$ — все остальные значения объекта, которые пропущены. Тогда в качестве оценки плотности объекта $x_{(k)}$ рассматриваем

$$\int\limits_{\mathbb{R}^{d_{(u)}}} p(x)dx_{(u)},$$

где $d_{(u)}$ — размерность вектора $x_{(u)}$.

Пусть объект $x \in \mathbb{R}^d$ признан аномальным. Исследователь может выбрать элементы вектора x, которым он доверяет, а другие оценить через них. Без ограничения общности считаем, что $x^T = (x_a^T, x_b^T)$ и значения x_b являются доверяемыми. Кроме того, по причине наличия пропусков в данных, некоторые значения x_a могут быть неизвестны.

Рассмотрим способ решения задачи регрессии. Вектор x_a при условии значений x_b имеет распределение смеси распределений с плотностью

$$\widetilde{p}(x) = \sum_{j=1}^{k} \widetilde{w}_j p(x|\widetilde{\theta}_j),$$

где $\widetilde{\theta}_j$ — параметры распределения j-й компоненты смеси при условии значений x_b. Подставив вместо параметров их оценки, получаем оценку условного распределения вектора x_a, которое на практике является достаточно информативным показателем для принятия различных выводов о векторе x_a. В частности, из него можно получить

- Оценку математического ожидания $\mathsf{E}(x_a \,|x_b)$. Эта оценка решает задачу регрессии признаков x_b на признаки x_a.
- Оценку дисперсии $\mathsf{D}(x_a \mid x_b)$.
- Оценку условных распределений кластеров всех объектов, у которых признаки X_b фиксированы и равны x_b.
- Множества наибольшей плотности — множество значений x_a, для которых плотность по условному распределению больше, чем для остальных значений.

Метод протестирован на данных PVT-свойств пластовых флюидов. База данных содержит результаты исследований более 3600 проб пластовых

флюидов. Среди рассматриваемых признаков имеются следующие величины: пластовое давление, температура пласта, поверхностная плотность газа, поверхностная плотность нефти, газосодержание, давление насыщения, пластовая плотность нефти, объемный коэффициент нефти, вязкость пластовой нефти.

Задача предсказания PVT-свойств методами машинного обучения ранее рассматривалась в сильно ограниченном варианте. Например, рассматривалось предсказание давление насыщения и объемного коэффициента нефти через другие свойства с помощью искусственных нейронных сетей и SVM-регрессии.

В результате данного исследования предложен принципиально другой подход к решению этих задач, основой которого является введение вероятностной модели в пространстве свойств пластовых флюидов, в которой предполагается, что признаковое описание пробы получается независимо от всех остальных проб из некоторого вероятностного распределения.

Для описания данных первоначально была применена смесь нормальных распределений из четырех компонент. В силу того, что нормальное распределение обладает легкими хвостами, оценки его параметров не устойчивы к наличию в данных шумовых объектов. Один из четырех кластеров описывал шумовые объекты, накрывая другие кластеры. Оценки других кластеров также смещены в сторону выбросов.

Для устранения перечисленных выше недостатков к данным применена модель смеси распределений Стьюдента из четырех компонент. Все четыре кластера соответствуют трем нешумовым кластером случая нормального распределения. Шумового кластера нет. Полученные кластеры нашли четкую интерпретацию у экспертов.

Проведенные эксперименты показали, что получаемые рекомендованные значения не противоречат физическим свойствам PVT-проб, в частности обладают гладкостью по аргументам. Кроме того, качество получаемого предсказания по метрикам MAPE и RMSPE в большинстве случаев превосходит другие известные модели машинного обучения.

Основные результаты диссертации. Получены следующее научные результаты, являющиеся новыми на период проведения исследований и опубликования:

1. Если $c = 1$, то $p_{b+1,n,c} > p_{b,n,c}$ при любых $1 \leqslant b < n$.

2. Если $n \geqslant 3b + 2$, то $p_{b+1,n,c} > p_{b,n,c}$.

3. Порог, при котором монотонность меняется, равен $\frac{n}{3(1-c)}(1+o(1))$. Формально, $\forall\varepsilon > 0\ \forall\delta > 0\ \forall c \in (0,1)\ \exists n_0\ \forall n \geqslant n_0\ \forall b \in (\varepsilon n, n)$:

   - при $b < \frac{n(1-\delta)}{3(1-c)}$ выполнено $p_{b+1,n,c} > p_{b,n,c}$,
   - при $b > \frac{n(1+\delta)}{3(1-c)}$ выполнено $p_{b+1,n,c} < p_{b,n,c}$.

4. Итерационная процедура получения оценки параметров в смеси многомерных распределений Стьюдента общего вида (без дополнительных ограничений на параметры) при помощи вариации EM-алгоритма, в которой на E-шаге применяется вариационный байесовский вывод. Данная процедура позволяет вычислять оценку параметров по выборке, в которых часть значений ненаблюдаема.

5. Метод машинного обучения, основанный на смеси многомерных распределений Стьюдента, позволяет решать следующие задачи:

   (a) кластеризация точек на выбранное количество кластеров,

   (b) выявление аномальных точек,

   (c) регрессия для предсказания любого набора вещественных признаков при использовании любого другого набора. Метод позволяет не только осуществлять предсказание, но и вычислять доверительный интервал.

   Каждая из этих задач может быть решена моделью при наличии пропусков в данных.

6. Программный продукт, реализующий вышеупомянутый метод, является новым инструментом для оценки достоверности данных исследований PVT-свойств пластовых флюидов.

Благодарности. Автор признателен доценту Максиму Евгеньевичу Жуковскому за неоценимую помощь в работе и за полезные замечания, Семену Андреевичу Буденному и Алле Михайловне Андриановой за помощь в постановке задач и экспертную поддержку.

Работы автора по теме диссертации

1. *Volkov, N., Dakhova, E., Budennyy, S., Andrianova, A.*. Student Mixture and Its Machine Learning Applications to PVT Properties of Reservoir Fluids // Advances in Systems Science and Applications. 2020. 20(2), 98-118. https://doi.org/10.25728/assa.2020.20.2.899

2. *Volkov, N., Andrianova, A., Serebryakova, D., Budennyy, S.*. Reliability Assessment of PVT-Properties of Reservoir Fluids on the Basis of a Probabilistic Mixture Model of Student's Distributions // Society of Petroleum Engineers. 2019. doi:10.2118/196866-MS

3. *Волков Н. А.* Монотонность функции биномиального распределения возле медианы // Труды МФТИ. 2020. V. 3(47). P. 3–16.

4. *Волков Н. А., Буденный С. А., Андрианова А. М.* Смеси вероятностных распределений в задачах регрессии и проверки на аномальность и их применение для PVT-свойств // Труды МФТИ. 2020. V. 3(47). P. 17–43.

5. *Andrianova, A., Simonov, M., Perets, D., Margarit, A., Serebryakova, D., Bogdanov, Y., Budennyy, S., Volkov, N., Tsanda, A., Bukharev, A.*. Application of Machine Learning for Oilfield Data Quality Improvement // Society of Petroleum Engineers. 2018. doi:10.2118/191601-18RPTC-MS