Байесовский подход к выбору достаточного размера выборки

Киселев Никита
kiselev.ns@phystech.edu

Грабовой Андрей
grabovoy.av@phystech.edu

15 февраля 2024 г.

Аннотация

Исследуется задача выбора достаточного размера выборки. Рассматривается проблема определения достаточного размера выборки без учета природы параметров используемой модели. Предлагается использовать функцию правдоподобия выборки. Рассматриваются два подхода к определению достаточного размера выборки через функцию правдоподобия. Предлагаемые подходы основываются на эвристиках о поведении функции правдоподобия при большом количестве объектов в выборке. Доказывается корректность предложенных подходов при определенных ограничениях на используемую модель. Предлагается метод прогнозирования функции правподобия в случае недостаточного размера выборки. Проводится вычислительный эксперимент для анализа свойств предложенных методов.

Ключевые слова: определение размера выборки · байесовский подход

1 Введение

Задача машинного обучения с учителем предполагает выбор предсказательной модели из некоторого параметрического семейства. Обычно такой выбор связан с некоторыми статистическими гипотезами, например, максимизацией некоторого функционала качества.

Определение 1. *Модель, которая соответствует этим статистическим гипотезам, называется* ***адекватной*** *моделью.*

При планировании вычислительного эксперимента требуется оценить минимальный размер выборки — количество объектов, необходимое для построения адекватной модели.

Определение 2. *Размер выборки, необходимый для построения адекватной модели прогнозирования, называется* ***достаточным****.*

В работе рассматривается проблема определения достаточного размера выборки. Этой теме посвящено большое число работ. Используемые в них подходы можно разделить на статистические, байесовские и эвристические.

Одни из первых статей по данной теме [1, 2] формулируют определенный статистический критерий, где связанный с данным критерием метод оценки размера выборки гарантирует достижение фиксированной статистической мощности с величиной ошибки первого рода, не превышающей заданного значения. Статистические методы имеют ряд ограничений, которые связаны с их применением на практике. Они позволяют оценить размер выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы.

Класс байесовских методов оценки размера выборки достаточно широк. В работе [3] достаточный размер выборки определяется исходя из максимизации ожидаемой функции полезности. Она может включать в себя в явном виде функции распределения параметров и штрафы за увеличение размера выборки. Также в этой работе рассматриваются альтернативные подходы, основанные на ограничении некоторого критерия качества оценки параметров модели. Среди таких критериев можно выделить критерий средней апостериорной дисперсии APVC, критерий среднего покрытия ACC, критерий срдней длины ALC и критерий эффективного объема выборки ESC. Эти критерии получили свое развитие в других работах, например, [4] и [5]. Спустя время, авторы [6] провели теоретическое и практическое сравнение методов из [1, 2, 3].

Авторы [7], как и [8], рассматривают различия между байесовским и частотным подходами при определении размера выборки. Также они предлагают робастные методы для байесовского подхода и приводят наглядные примеры для некоторых вероятностных моделей.

В работе [9] рассматриваются различные методы оценки размера выборки в обобщенных линейных моделях, включая статистические, эвристические и байесовские методы. Анализируются такие методы, как тест на множители Лагранжа, тест на отношение правдоподобия, статистика Вальда, кросс-валидация, бутстрап, критерий Кульбака-Лейблера, критерий средней апостериорной дисперсии, критерий среднего охвата, критерий средней длины и максимизация полезности. Авторы указывают на возможное развитие темы, которое заключается в поиске метода, сочетающего байесовский и статистический подходы для оценки размера выборки для недостаточного доступного размера выборки.

В [10] рассматривается метод определения размера выборки в логистической регрессии, использующий кросс-валидацию и дивергенцию Кульбака-Лейблера между апостериорными распределениями параметров модели на схожих подвыборках. Под схожими подвыборками понимают такие подвыборки, которые могут быть получены друг из друга добавлением, удалением или заменой одного объекта.

Настоящая работа использует генетический алгоритм [11] с целью аппроксимации заданного набора функций. Генетический алгоритм представляет собой процесс оптимизации популяции кандидатов (называемых особями), который эволюционирует в сторону лучших решений [12]. Каждая особь имеет набор характеристик (генов или фенотипов), которые могут изменяться в процессе эволюции. Изменение происходит

с помощью операции кроссинговера или мутации. Эволюция начинается с случайной популяции, и каждое поколение рассматривается как основа для генерации следующего. Приспособленность особей измеряется в каждом поколении, и особи с высокой приспособленностью выбираются для создания нового поколения [13]. Алгоритм завершается после достижения максимального числа поколений или достижения удовлетворительных результатов. Таким образом, каждое новое поколение становится более приспособленным к окружающей среде.

2 Постановка задачи

Объектом называется пара $(\mathbf{x}, y)$, где $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ есть вектор признакового описания объекта, а $y \in \mathbb{Y}$ есть значение целевой переменной. В задаче регрессии $\mathbb{Y} = \mathbb{R}$, а в задаче K-классовой классификации $\mathbb{Y} = \{1, \ldots, K\}$.

Матрицей объекты-признаки для выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется матрица $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$.

Вектором ответов (вектором значений целевой переменной) для выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется вектор $\mathbf{y} = [y_1, \ldots, y_m]^\intercal \in \mathbb{Y}^m$.

Моделью называется параметрическое семейство функций f, отображающих декартово произведение множества значений признакового описания объектов $\mathbb{X}$ и множества значений параметров $\mathbb{W}$ во множество значений целевой переменной $\mathbb{Y}$:

$$f : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}.$$

Вероятностной моделью называется совместное распределение вида

$$p(y, \mathbf{w}|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}) : \mathbb{Y} \times \mathbb{W} \times \mathbb{X} \to \mathbb{R}^+,$$

где $\mathbf{w} \in \mathbb{W}$ есть набор параметров модели, $p(y|\mathbf{x}, \mathbf{w})$ задает правдоподобие объекта, а $p(\mathbf{w})$ задает априорное распределение параметров.

Функцией правдоподобия простой выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется функция

$$L(\mathfrak{D}_m, \mathbf{w}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Ее логарифм

$$l(\mathfrak{D}_m, \mathbf{w}) = \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \mathbf{w})$$

называется логарифмической функцией правдоподобия. Далее, если не оговорено противное, будем считать выборку простой.

Оценкой максимума правдоподобия набора параметров $\mathbf{w} \in \mathbb{W}$ по выборке $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется

$$\hat{\mathbf{w}}_m = \arg\max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_m, \mathbf{w}).$$

Ставится задача определения достаточного размера выборки m^*. Пусть задан некоторый критерий T. Он может быть построен, например, на основе эвристик о поведении параметров модели.

Определение 3. *Размер выборки m^* называется **достаточным** согласно критерию T, если T выполняется для всех $k \geqslant m^*$.*

Стоит учесть, что возможно $m^* \leqslant m$ или $m^* > m$. Эти два случая будут отдельно рассмотрены далее.

3 Достаточный размер выборки не превосходит доступный

В этом разделе будем считать, что достоверно $m^* \leqslant m$. Это означает, что нам нужно просто формализовать, какой размер выборки можно считать достаточным.

3.1 Анализ поведения функции правдоподобия

Для определения достаточности будем использовать функцию правдоподобия. Когда в наличии имеется достаточно объектов, вполне естественно ожидать, что от одной реализации выборки к другой полученная оценка параметров не будет сильно меняться [2, 14]. То же можно сказать и про функцию правдоподобия. Таким образом, сформулируем, какой размер выборки можно считать достаточным.

Определение 4. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **D-достаточным**, если для всех $k \geqslant m^*$*

$$D(k) = \mathbb{D}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leqslant \varepsilon.$$

С другой стороны, когда в наличии имеется достаточно объектов, также вполне естественно, что при добавлении очередного объекта в рассмотрение полученная оценка параметров не будет сильно меняться. Сформулируем еще одно определение.

Определение 5. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **M-достаточным**, если для всех $k \geqslant m^*$*

$$M(k) = \left|\mathbb{E}_{\mathfrak{D}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)\right| \leqslant \varepsilon.$$

Замечание. *В определениях 4 и 5 вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ можно рассматривать ее логарифм $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.*

Предположим, что $\mathbb{W} = \mathbb{R}^n$, т.е. параметры $\mathbf{w}$ представляются в виде вектора. Напомним, что информацией Фишера называется матрица

$$[\mathcal{I}(\mathbf{w})]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\partial w_i \partial w_j}\right].$$

Достаточно известным результатом является асимптотическая нормальность оценки максимума правдоподобия.

Утверждение 1. *Пусть* $\hat{\mathbf{w}}_k$ *есть оценка максимума правдоподобия параметров* $\mathbf{w}$. *Тогда при определенных условиях регулярности (которые на практике чаще всего выполнены) имеет место следующая сходимость по распределению:*

$$\sqrt{k}\left(\hat{\mathbf{w}}_k - \mathbf{w}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\mathbf{w})\right),$$

откуда следует

$$\hat{\mathbf{w}}_k \xrightarrow{d} \mathcal{N}\left(\mathbf{w}, \left[k\mathcal{I}(\mathbf{w})\right]^{-1}\right).$$

Из сходимости по распределению в общем случае не следует сходимость моментов случайного вектора. Тем не менее, если предположить последнее, то в некоторых моделях можно доказать корректность предложенного нами определения M-достаточного размера выборки.

Для удобства обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом: математическое ожидание $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и матрица ковариации $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$. Тогда имеет место следующая теорема, доказательство которой приведено в разделе 7.

Теорема 1 (Киселев, 2023)**.** *Пусть* $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ *и* $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ *при* $k \to \infty$. *Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным. А именно, для любого* $\varepsilon > 0$ *найдется такой* m^*, *что для всех* $k \geqslant m^*$ *выполнено* $M(k) \leqslant \varepsilon$.

Следствие. *Пусть* $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ *и* $\|\mathbf{\Sigma}_k - \left[k\mathcal{I}(\mathbf{w})\right]^{-1}\|_F \to 0$ *при* $k \to \infty$. *Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным.*

По условию задана одна выборка. Поэтому в эксперименте нет возможности посчитать указанные в определениях математическое ожидание и дисперсию. Для их оценки воспользуемся техникой бутстрап. А именно, сгенерируем из заданной $\mathfrak{D}_m$ некоторое число B подвыборок размера k с возвращением. Для каждой из них получим оценку параметров $\hat{\mathbf{w}}_k$ и посчитаем значение $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Для оценки будем использовать выборочное среднее и несмещенную выборочную дисперсию (по бутстрап-выборкам).

Предложенные выше определения можно применять и в тех задачах, когда минимизируется произвольная функция потерь, а не максимизируется функция правдоподобия. Мы не приводим никаких теоретических обоснований этого, однако на практике такая эвристика оказывается достаточно удачной.

3.2 Анализ апостериорного распределения параметров модели

В работе [10] предлагается использовать дивергенцию Кульбака-Лейблера для оценки достаточного размера выборки в задаче бинарной классификации. Идея основывается на том, что если две подвыборки отличаются друг от друга на один объект, то полученные по ним апостериорные распределения должны быть близки. Эта близость определяется дивергенцией Кульбака-Лейблера.

В настоящей работе предлагается развить этот подход, исследовать его не только в задаче классификации, но и в задаче регрессии. В качестве меры близости предлагается

использовать не только дивергенцию Кульбака-Лейблера, но и функцию сходства s-score из [Адуенко].

Рассмотрим две подвыборки $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ и $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Пусть $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \ldots, m\}$ и $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \ldots, m\}$ — соответствующие им подмножества индексов.

Определение 6. *Подвыборки $\mathfrak{D}^1$ и $\mathfrak{D}^2$ называются **схожими**, если $\mathcal{I}_2$ может быть получено из $\mathcal{I}_1$ удалением, заменой или добавлением одного элемента, то есть*

$$|\mathcal{I}_1 \triangle \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Рассмотрим две схожие подвыборки $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ и $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ размеров k и $k+1$ соответственно. Это означает, что большая из них получена добавлением одного элемента к меньшей. Найдем апостериорное распределение параметров модели по этим подвыборкам:

$$p_k(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_k) = \frac{p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_k)} \propto p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w}),$$

$$p_{k+1}(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_{k+1}) = \frac{p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_{k+1})} \propto p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w}).$$

Определение 7. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **KL-достаточным**, если для всех $k \geqslant m^*$*

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leqslant \varepsilon.$$

Для пары нормальных распределений дивергенция Кульбака-Лейблера имеет достаточно простой вид. Предположим, что апостериорное распределение является нормальным, то есть $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \mathbf{\Sigma}_k)$. Руководствуясь эвристикой, что сходимость моментов такого распределения должна влечь за собой близость апостериорных распределений на схожих подвыборках, можно сформулировать следующее утверждение.

Теорема 2 (Киселев, 2024)**.** *Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ при $k \to \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение KL-достаточного размера выборки является корректным. А именно, для любого $\varepsilon > 0$ найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $KL(k) \leqslant \varepsilon$.*

В настоящей работе предлагается в качестве меры сходства распределений использовать меру сходства s-score из [Адуенко]:

$$\text{s-score}(p_1, p_2) = \frac{\int_{\mathbf{w}} p_1(\mathbf{w}) p_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} p_1(\mathbf{w} - \mathbf{b}) p_2(\mathbf{w}) d\mathbf{w}}.$$

Определение 8. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **S-достаточным**, если для всех $k \geqslant m^*$*

$$S(k) = s\text{-}score(p_k, p_{k+1}) \geqslant 1 - \varepsilon.$$

Как и в случае KL-достаточного размера выборки, в модели с нормальным апостериорным распределением есть возможность записать выражение для используемого критерия. Таким образом, можно сформулировать еще одно утверждение.

Теорема 3 (Киселев, 2024). *Пусть* $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ *при* $k \to \infty$. *Тогда в модели с нормальным апостериорным распределением параметров определение S-достаточного размера выборки является корректным. А именно, для любого* $\varepsilon > 0$ *найдется такой* m^*, *что для всех* $k \geqslant m^*$ *выполнено* $S(k) \geqslant 1 - \varepsilon$.

Пусть в модели линейной регрессии задано нормальное априорное распределение параметров. По свойству сопряженности априорного распределения и правдоподобия апостериорное распределение также является нормальным. Таким образом, мы приходим к одному из простейших примеров модели, для которой справедливы теоремы, представленные выше. На самом деле для линейной регрессии можно сформулировать более простые утверждения.

Теорема 4 (Киселев, 2024). *Пусть* $\left\|\mathbf{X}_{k+1}^\intercal \mathbf{X}_{k+1} - \mathbf{X}_k^\intercal \mathbf{X}_k\right\|_2 = o(k^{-1/2})$ *и* $\lambda_{\min}\left(\mathbf{X}_k^\intercal \mathbf{X}_k\right) = \Omega(1)$ *при* $k \to \infty$. *Тогда в модели линейной регрессии с нормальным априорным распределением параметров* $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ *при* $k \to \infty$.

Теорема 5 (Киселев, 2024). *Пусть... Тогда в модели линейной регрессии с нормальным априорным распределением параметров* $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ *при* $k \to \infty$.

4 Достаточный размер выборки больше доступного

В этом разделе будем считать, что достоверно $m^* > m$.

Возникает задача прогнозирования математического ожидания функции правдоподобия / функции ошибки при $k > m$. В общем виде это достаточно трудная задача. В настоящей работе предлагается проанализировать большое число открытых датасетов из [15], чтобы найти параметрическое семейство функций, которыми стоит аппроксимировать зависимость функции ошибки от используемого размера выборки. Предлагается отдельно исследовать датасеты с задачами регрессии и классификации.

4.1 Генетический алгоритм в задаче аппроксимации набора функций

Одним из наиболее простых с точки зрения реализации и логики алгоритмов перебора является генетический алгоритм. С помощью него построим метод нахождения искомого семейства функций.

Пусть для N различных датасетов построен график зависимости среднего значения функции ошибки (или функции правдоподобия со знаком минус) от используемого размера выборки. Приведем эти N зависимостей к одинаковому масштабу по обеим осям. Для этого вычтем минимальное значение, а затем поделим на максимальное значение. В таком случае график каждой зависимости лежит в квадрате $[0; 1]^2$, начинается в точке $(0; 1)$ и заканчивается в точке $(1; 0)$.

Популяцией в генетическом алгоритме является набор параметрических семейств функций. Например, одной особью может является семейство $w_0 + w_1 \cdot \log(w_2 \cdot x) +$

$w_3 \cdot x^2$, где x есть переменная, а $\mathbf{w}$ есть вектор параметров. Начальная популяция инициализируется случайным образом. Используются простейшие унарные функции: $1, x, \sin x, \cos x, \exp x, \log x, \operatorname{ctg} x$ и $\operatorname{cth} x$, а также простейшие бинарные функции: $+, -, *$ и $/$. Каждая особь представляется с помощью бинарного дерева, в узлах которого стоят вышеупомянутые функции, а листьями являются обязательно 1 или x. При этом за каждым узлом закрепляется своя компонента вектора параметров.

Приспособленность особи измеряется следующим образом. Для каждой из N аппроксимируемых зависимостей решается задача подбора вектора параметров. Минимизируется среднеквадратичное отклонение. Полученное значение MSE усредняется по всем N зависимостям. Итоговое значение определяет приспособленность особи.

Кроссинговер реализуется так, что случайное поддерево одного из особей-родителей заменяется случайным поддеревом другого. Мутация заменяет функцию в случайном узле дерева на другую случайную функцию.

Алгоритм завершается по прошествии заданного числа поколений. Выбирается особь из последнего поколения с наилучшей приспособленностью. Решением является соответствующее параметрическое семейство функций.

5 Вычислительный эксперимент

Проводится эксперимент для анализа свойств предложенных методов оценки достаточного размера выборки. Эксперимент состоит из нескольких частей. В первой части рассматриваются оценки достаточного размера выборки в случае, когда достаточный размер выборки не превосходит доступный. Во второй части исследуются результаты, полученные в условиях того, что достаточный размер выборки больше доступного.

5.1 Достаточный размер выборки не превосходит доступный

Синтетические данные сгенерированы из модели линейной регрессии. Число объектов 1000, число признаков 20. Далее приведены графики логарифма функции правдоподобия выборки, а также функций $D(k)$ и $M(k)$, определенных в Главе 3 (здесь используется логарифм функции правдоподобия). Выполнено определение D-достаточного и M-достаточного размеров выборки. Использовалось $B = 1000$ бутстрап-выборок. Результаты представлены на Рис. 1.

Вторая синтетическая выборка сгенерирована из модели логистической регрессии. Число объектов 1000, число признаков 20. Аналогичные графики приведены на Рис. 2.

5.2 Достаточный размер выборки больше доступного

5.2.1 Определение параметрического семейства функций с помощью генетического алгоритма

Реализацию генетического алгоритма, приведенного в разделе 4.1, можно найти в репозитории. Для исследования зависимости функции ошибки от используемого размера выборки в задаче регрессии использовались следующие датасеты из [15]: Abalone, Auto

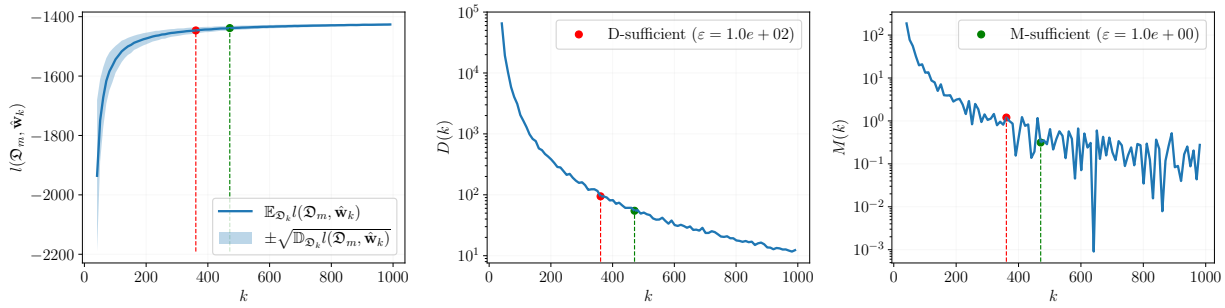

Рис. 1: Синтетическая выборка (линейная регрессия) при $m^* \leqslant m$

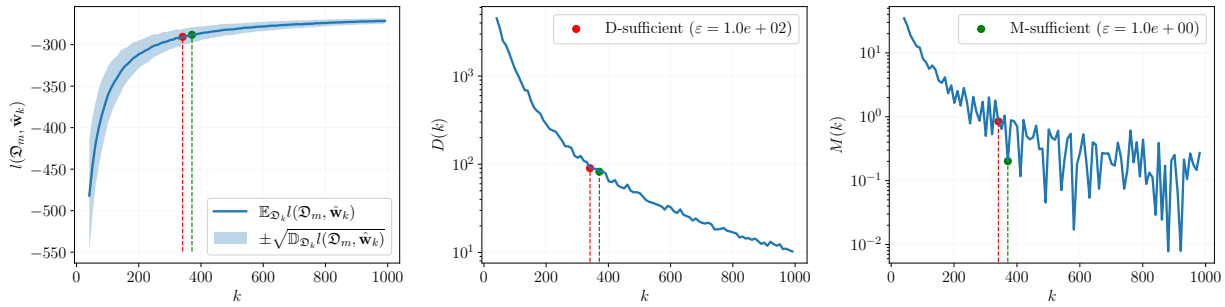

Рис. 2: Синтетическая выборка (логистическая регрессия) при $m^* \leqslant m$

MPG, Liver Disorders, Wine Quality, Parkinsons Telemonitoring, Bike Sharing Dataset, Real estate valuation и Heart failure clinical records. Была выбрана квадратичная функция потерь MSE. Задача регрессии для каждого из них решалась с помощью линейной регрессии из [16]. Усреднение производилось по $B = 100$ бутстрап-выборкам. Как было сказано ранее, все зависимости приводятся к одинаковому масштабу по обеим осям. Полученные графики представлены на Рис. 3. Слева находится график для выборочного среднего. Справа находится график для выборочного среднеквадратичного отклонения.

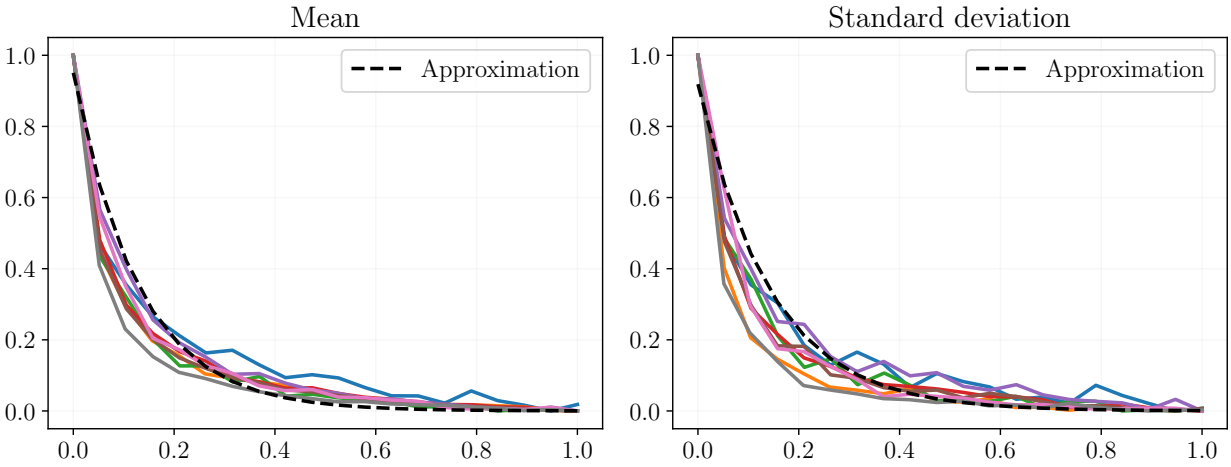

Рис. 3: Поведение функции ошибки в задаче регрессии

Применение генетического алгоритма приводит к одинаковому семейству функций для аппроксимации среднего и среднеквадратичного отклонения в задаче регрессии:

$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

В задаче классификации использовалось 12 датасетов из [15]: Automobile, Breast Cancer Wisconsin (Diagnostic), Car Evaluation, Credit Approval, Glass Identification, Ionosphere, Iris, Tic-Tac-Toe Endgame, Congressional Voting Records, Wine, Zoo и Heart failure clinical records. Задача классификации для каждого из них решалась с помощью логистической регрессии из [16]. Усреднение производилось по $B = 100$ бутстрап-выборкам. Все завимисости также приводятся к одинаковому масштабу по обеим осям. Полученные графики представлены на Рис. 4. Как и ранее, слева находится график для выборочного среднего, справа находится график для выборочного среднеквадратичного отклонения.

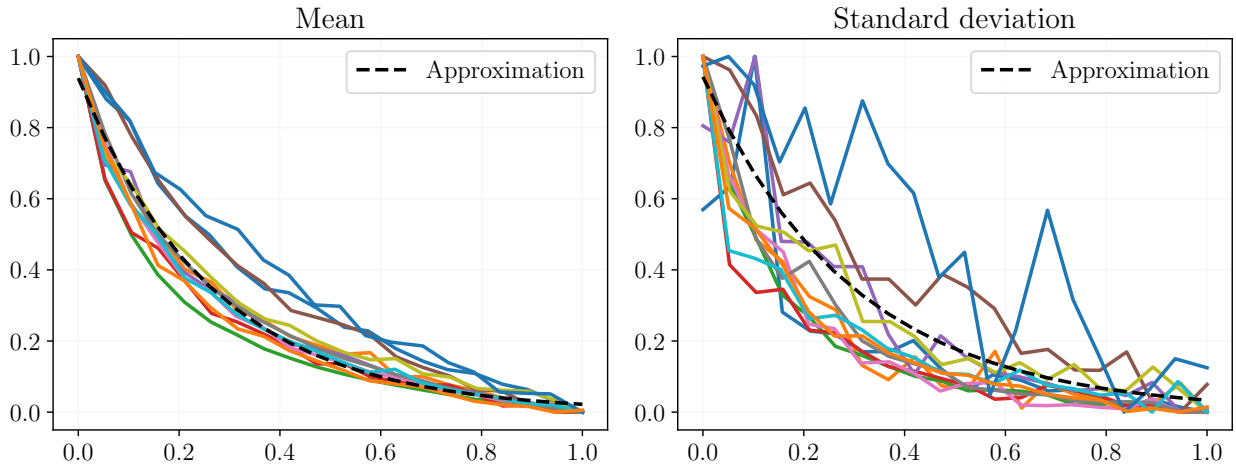

Рис. 4: Поведение функции ошибки в задаче классификации

Применение генетического алгоритма для среднего значения приводит к такому же семейству функций, как и в задаче регрессии:

$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

Среднеквадратичное отклонение в случае задачи классификации для каждой выборки имеет свою зависимость от размера выборки. Таким образом, прогнозировать дисперсию для классификации оказывается достаточно сложной задачей.

5.2.2 Прогнозирование функции правдоподобия

Для синтетических выборок проведена аппроксимация функций правдоподобия. Среднее значение и дисперсия аппроксимированы параметрическим семейством функций, приведенным в предыдущем пункте.

Производилось разделение на обучающую и тестовую выборки в соотношении 70:30. Аппроксимация производилась только на обучающей части. Достаточный размер выбор-

ки находился в тестовой части. На Рис. 5 и Рис. 6 представлены истинные и восстановленные зависимости. Там же указаны определенные D-достаточный и M-достаточный размеры выборки.

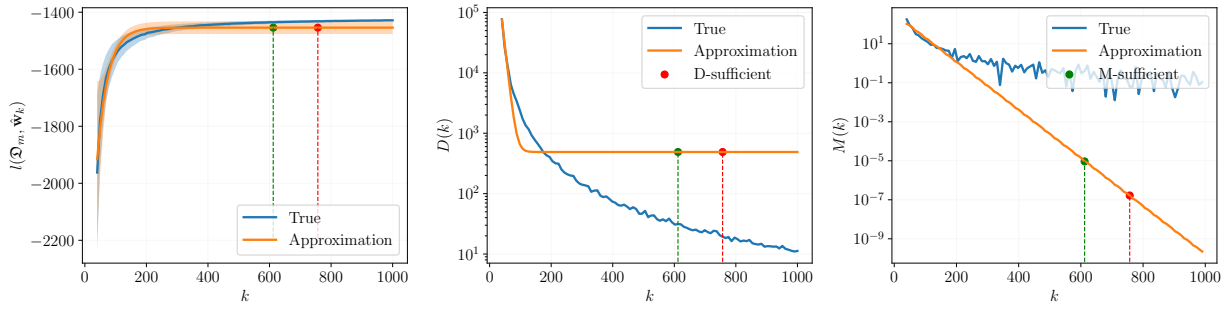

Рис. 5: Синтетическая выборка (линейная регрессия) при $m^* > m$

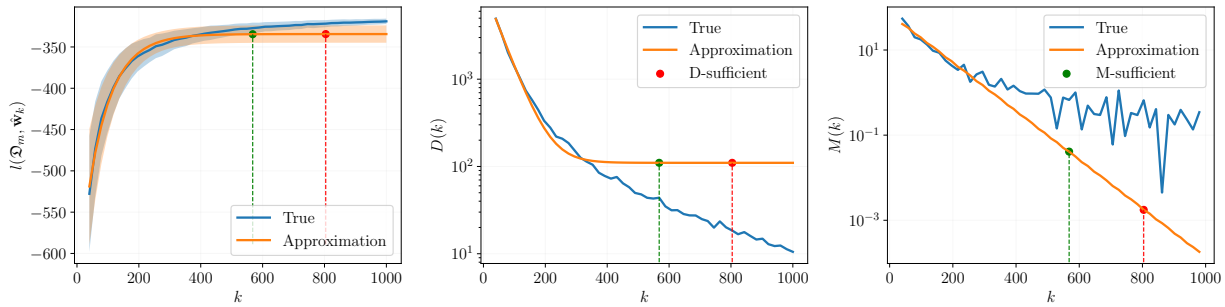

Рис. 6: Синтетическая выборка (логистическая регрессия) при $m^* > m$

6 Заключение

Основные результаты данной работы заключаются в следующем. Предложены способы определения достаточного размера выборки, не использующие распределение параметров модели. Это позволяет применять их для любой модели прогнозирования, будь то линейная регрессия или нейронная сеть. В частном случае доказана корректность такого определения. Проведены вычислительные эксперименты, подтверждающие теоретические выкладки и демонстрирующие работу метода.

Список литературы

[1] C. J. Adcock. A bayesian approach to calculating sample sizes. *The Statistician*, 37 (4/5):433, 1988. ISSN 0039-0526. doi:10.2307/2348770. URL http://dx.doi.org/10.2307/2348770.

[2] Lawrence Joseph, David B. Wolfson, and Roxane Du Berger. Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(2):143–154, 1995. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2348439.

[3] Dennis V. Lindley. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):129–138, July 1997. ISSN 1467-9884. doi:10.1111/1467-9884.00068. URL http://dx.doi.org/10.1111/1467-9884.00068.

[4] T. Pham-Gia. On bayesian analysis, bayesian decision theory and the sample size problem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46 (2):139–144, July 1997. ISSN 1467-9884. doi:10.1111/1467-9884.00069. URL http://dx.doi.org/10.1111/1467-9884.00069.

[5] Alan E. Gelfand and Fei Wang. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2), May 2002. ISSN 0883-4237. doi:10.1214/ss/1030550861. URL http://dx.doi.org/10.1214/ss/1030550861.

[6] Jing Cao, J. Jack Lee, and Susan Alber. Comparison of bayesian sample size criteria: Acc, alc, and woc. *Journal of Statistical Planning and Inference*, 139(12):4111–4122, December 2009. ISSN 0378-3758. doi:10.1016/j.jspi.2009.05.041. URL http://dx.doi.org/10.1016/j.jspi.2009.05.041.

[7] Pierpaolo Brutti, Fulvio De Santis, and Stefania Gubbiotti. Bayesian-frequentist sample size determination: a game of two priors. *METRON*, 72(2):133–151, May 2014. ISSN 2281-695X. doi:10.1007/s40300-014-0043-2. URL http://dx.doi.org/10.1007/s40300-014-0043-2.

[8] Hamid Pezeshk, Nader Nematollahi, Vahed Maroufy, and John Gittins. The choice of sample size: a mixed bayesian / frequentist approach. *Statistical Methods in Medical Research*, 18(2):183–194, April 2008. ISSN 1477-0334. doi:10.1177/0962280208089298. URL http://dx.doi.org/10.1177/0962280208089298.

[9] A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, and V. V. Strijov. Numerical methods of sufficient sample size estimation for generalised linear models. *Lobachevskii Journal of Mathematics*, 43(9):2453–2462, September 2022. ISSN 1818-9962. doi:10.1134/s1995080222120125. URL http://dx.doi.org/10.1134/S1995080222120125.

[10] Anastasiya Motrenko, Vadim Strijov, and Gerhard-Wilhelm Weber. Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255:743–752, 2014. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2013.06.031. URL https://www.sciencedirect.com/science/article/pii/S0377042713003294.

[11] David E. Goldberg and John H. Holland. Genetic algorithms and machine learning. *Machine Learning*, 3(2):95–99, 1988. doi:10.1023/A:1022602019183. URL https://doi.org/10.1023/A:1022602019183.

[12] Seyedali Mirjalili. *Genetic Algorithm*, pages 43–55. Springer International Publishing, Cham, 2019. ISBN 978-3-319-93025-1. doi:10.1007/978-3-319-93025-1_4. URL https://doi.org/10.1007/978-3-319-93025-1_4.

[13] Oliver Kramer. *Genetic Algorithms*, pages 11–19. Springer International Publishing, Cham, 2017. ISBN 978-3-319-52156-5. doi:10.1007/978-3-319-52156-5_2. URL https://doi.org/10.1007/978-3-319-52156-5_2.

[14] Lawrence Joseph, Roxane Du Berger, and Patrick Bélisle. Bayesian and mixed bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16(7): 769–781, 1997. doi:https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7<769::AID-SIM495>3.0.CO;2-V. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819970415%2916%3A7%3C769%3A%3AAID-SIM495%3E3.0.CO%3B2-V.

[15] Kolby Nottingham Markelle Kelly, Rachel Longjohn. The uci machine learning repository. URL https://archive.ics.uci.edu.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

7 Приложение

Доказательство (Теорема 1). Рассмотрим определение М-достаточного размера выборки в терминах логарифма функции правдоподобия. В модели линейной регрессии

$$L\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} \mathcal{N}\left(y_i|\hat{\mathbf{w}}_k^{\intercal}\mathbf{x}_i, \sigma^2\right) =$$

$$= \left(2\pi\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right).$$

Прологарифмируем:

$$l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2.$$

Возьмем математическое ожидание по $\mathfrak{D}_k$, учитывая, что $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и $\operatorname{cov}(\hat{\mathbf{w}}_k) = \boldsymbol{\Sigma}_k$:

$$\mathbb{E}_{\mathfrak{D}_k} l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 + \operatorname{tr}\left(\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\Sigma}_k\right)\Big).$$

Запишем выражение для разности математических ожиданий:

$$\mathbb{E}_{\mathfrak{D}_{k+1}} l(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) =$$

$$= \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 - \|\mathbf{y} - \mathbf{X}\mathbf{m}_{k+1}\|_2^2\Big) + \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^{\intercal}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big) =$$

$$= \frac{1}{2\sigma^2}\Big(2\mathbf{y}^{\intercal}\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\Big) +$$

$$+ \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^{\intercal}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big).$$

Значение функции $M(k)$ есть модуль от вышеприведенного выражения. Применим неравенство треугольника для модуля, а затем оценим каждое слагаемое.
Первое слагаемое оценим, используя неравенство Коши-Буняковского:

$$\left|\mathbf{y}^{\intercal}\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{X}^{\intercal}\mathbf{y}\|_2\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2.$$

Второе слагаемое оценим, используя неравенство Коши-Буняковского, свойство согласованности спектральной матричной нормы, а также ограниченность последовательности векторов $\mathbf{m}_k$, которая следует из предъявленной в условии сходимости:

$$\left|(\mathbf{m}_k - \mathbf{m}_{k+1})^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\right| \leqslant \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2\|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leqslant$$

$$\leqslant \|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2\|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leqslant C\|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.$$

Последнее слагаемое оценим, используя неравенство Гельдера для нормы Фробениуса:

$$\left|\operatorname{tr}\Big(\mathbf{X}^{\intercal}\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big)\right| \leqslant \|\mathbf{X}^{\intercal}\mathbf{X}\|_F\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F.$$

Наконец, поскольку $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$, то $M(k) \to 0$ при $k \to \infty$, что доказывает теорему. □

Доказательство (Следствие). Из приведенных в условии сходимостей следует, что $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$. Применение Теоремы 1 заканчивает доказательство. □

Доказательство (Теорема 2). Дивергенция Кульбака-Лейблера для пары нормальных апостериорных распределений имеет вид

$$D_{\mathrm{KL}}\left(p_k \| p_{k+1}\right) = \frac{1}{2}\left(\mathrm{tr}\left(\mathbf{\Sigma}_{k+1}^{-1}\mathbf{\Sigma}_k\right) + (\mathbf{m}_{k+1} - \mathbf{m}_k)^{\intercal}\mathbf{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k) - n + \log\left(\frac{\det \mathbf{\Sigma}_{k+1}}{\det \mathbf{\Sigma}_k}\right)\right).$$

Представим $\mathbf{\Sigma}_{k+1}$ как $\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}$. Рассмотрим в отдельности каждое слагаемое.

$$\mathrm{tr}\left(\mathbf{\Sigma}_{k+1}^{-1}\mathbf{\Sigma}_k\right) = \mathrm{tr}\left(\left(\mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}\right)^{-1}\mathbf{\Sigma}_k\right) \to \mathrm{tr}\mathbf{I}_n = n \text{ при } \|\Delta\mathbf{\Sigma}\|_F \to 0,$$

$$\left|(\mathbf{m}_{k+1} - \mathbf{m}_k)^{\intercal}\mathbf{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2\|\mathbf{\Sigma}_{k+1}^{-1}\|_2 \to 0 \text{ при } \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0,$$

$$\log\left(\frac{\det \mathbf{\Sigma}_{k+1}}{\det \mathbf{\Sigma}_k}\right) = \log\left(\frac{\det\left(\mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}\right)}{\det \mathbf{\Sigma}_k}\right) \to \log\det \mathbf{I}_n = \log 1 = 0 \text{ при } \|\Delta\mathbf{\Sigma}\|_F \to 0,$$

откуда и имеем требуемое. □

Доказательство (Теорема 3). Воспользуемся выражением s-score для пары нормальных априорных распределений из [Адуенко]:

$$\text{s-score}(p_k, p_{k+1}) = \exp\left(-\frac{1}{2}(\mathbf{m}_{k+1} - \mathbf{m}_k)^{\intercal}\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}\left(\mathbf{m}_{k+1} - \mathbf{m}_k\right)\right).$$

Поскольку

$$\left|(\mathbf{m}_{k+1} - \mathbf{m}_k)^{\intercal}\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}\left(\mathbf{m}_{k+1} - \mathbf{m}_k\right)\right| \leqslant \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2\|\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}\|_2 \to 0$$

при $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$, то значение квадратичной формы внутри экспоненты стремится к нулю. Следовательно, s-score$(p_k, p_{k+1}) \to 1$ при $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$. □

Доказательство (Теорема 4). Пусть задано нормальное априорное распределение параметров $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \tau^{-1}\mathbf{I}\right)$. В модели линейной регрессии правдоподобие является нормальным, а именно

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right) = \left(2\pi\sigma^2\right)^{-m/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2\right).$$

Используя сопряженность априорного распределения и правдоподобия, легко найти параметры апостериорного распределения:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma}\right),$$

где

$$\mathbf{\Sigma} = \left(\tau\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}, \qquad \mathbf{m} = \left(\mathbf{X}^{\intercal}\mathbf{X} + \tau\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{y}.$$

Рассмотрим выражение $\|\mathbf{\Sigma}_{k+1}-\mathbf{\Sigma}_k\|_2$ нормы разности матриц ковариции для подвыборок размера k и $k+1$. Введем обозначение $\mathbf{A}_k = \frac{1}{\sigma^2}\mathbf{X}_k^\intercal\mathbf{X}_k$. Учитывая формулы выше, имеем

$$\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_2 = \left\|(\tau\mathbf{I} + \mathbf{A}_{k+1})^{-1} - (\tau\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 =$$

$$= \left\|(\tau\mathbf{I} + \mathbf{A}_{k+1})^{-1}(\mathbf{A}_{k+1} - \mathbf{A}_k)(\tau\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 \leqslant$$

Воспользуемся субмультипликативностью спектральной матричной нормыж

$$\leqslant \left\|(\tau\mathbf{I} + \mathbf{A}_{k+1})^{-1}\right\|_2 \left\|(\tau\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 \|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 =$$

Теперь воспользуемся выражением спектральной матричной нормы через максимальное собственное значение.

$$= \frac{1}{\lambda_{\min}(\tau\mathbf{I} + \mathbf{A}_{k+1})}\frac{1}{\lambda_{\min}(\tau\mathbf{I} + \mathbf{A}_k)}\|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 \leqslant$$

$$\leqslant \frac{1}{\lambda_{\min}(\mathbf{A}_{k+1})}\frac{1}{\lambda_{\min}(\mathbf{A}_k)}\|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 =$$

$$= \sigma^2\frac{1}{\lambda_{\min}\left(\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1}\right)}\frac{1}{\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k)}\left\|\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} - \mathbf{X}_k^\intercal\mathbf{X}_k\right\|_2.$$

Поскольку по условию $\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k) = \Omega(1)$ при $k \to \infty$, то произведение дробей в выражении выше ограничено константой, начиная с некоторого номера. Так как, кроме того, $\left\|\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} - \mathbf{X}_k^\intercal\mathbf{X}_k\right\|_2 = o(k^{-1/2})$ при $k \to \infty$, то и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_2 = o(k^{-1/2})$ при $k \to \infty$. Далее воспользуемся эквивалентностью матричных норм, а именно

$$\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \leqslant \sqrt{k}\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_2 = o(1) \text{ при } k \to \infty,$$

что и требовалось доказать. □

Доказательство (Теорема 5). В обозначениях предыдущей теоремы имеем

$$\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 = \left\|\left(\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} + \tau\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_{k+1}^\intercal\mathbf{y}_{k+1} - \left(\mathbf{X}_k^\intercal\mathbf{X}_k + \tau\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_k^\intercal\mathbf{y}_k\right\|_2 \leqslant$$

□