Содержание

Аннотация

Исследована проблема снижения затрат на сбор данных, необходимых для построения адекватной модели. Рассматриваются задачи линейной и логистической регрессий. Для решения этих задач требуется, чтобы выборка содержала достаточное число объектов. Требуется предложить метод вычисления оптимального обьема данных, соблюдая при этом баланс между точностью модели и и трудозатратами при сборе данных. Предпочтительны те методы оценки объема, которые позволяют строить адекватные модели по выборкам возможно меньшего объема.

Ключевые слова: *планирование эксперимента, обобщенная линейная модель, оценка объёма выборки, прогнозирование объёма выборки, линейная регрессия, логичтическая регрессия, бутстреп.*

1. Введение

При планировании эксперимента требуется оценить минимальный достаточный объём выборки — число производимых измерений набора показателей или признаков, необходимый для построениия модели с оптимальным качеством.

Существует большое количество методов оценки достаточного объёма выборки. Например, тест множителей Лагранжа, тест отношения правдоподобия и тест Вальда. В работах [?, ?, ?] на основе данных методов построена оценка достаточного объёма выборки. Основной минус этих методов заключается в том, что статистики, используемые в критериях, имеют асимптотическое распределение и требуют большого объёма выборки.

Существуют байесовские оценки обьема выборки: критерий средней апостериорной дисперсии, критерий среднего покрытия, критерий средней длины и метод максимизации полезности. Первые три метода, описанные в работе [?], требуют анализа некоторой функции эффективности от размера выборки. Используя некоторое решающее правило, по данной функции определяется достаточный объем выборки. Главный минус этих методов заключается в том, что они не позволяют построить аппроксимацию функции эффективности при большем объеме данных. Метод максимизации полезности максимизирует ожидание некоторой функции полезности по объёму выборки. Все эти методы опираются на апостериорное распределение, что требует достаточно большого обьема выборки.

В работе [?] описан вывод функций ошибки для задач линейной и логистической регрессий для обобщённых линейных моделей. Функция ошибки назначается путём байесовского вывода и определяется гипотезой порождения данных.

Существуют также модели, аппроксимирующие зависимость функции ошибки от объёма выборки и предсказывающие достаточный объём выборки. В работах [?, ?] описаны такие модели для решения определённой задачи классификации на конкретном наборе данных. В работе [?] из датасета извлекаются характерные для задачи признаки (дисперсия, количество различных генов и

т.д.), которые далее используются для построения аппроксимирующей модели.

Решается задача прогнозирования достаточного объёма выборки на раннем этапе сбора данных. Предполагается, что зависимая переменная аппроксимируется обобщенной линейной моделью. Требуется определить оптимальный набор признаков.

Для построения модели необходимо решить задачу отбора значимых признаков. Существует множество методов для решения этой задачи: Лассо [**?**], Stagewise [**?**], LARS [**?**], Random Forests [**?**]. В данной работе предлагается использовать метод отбора признаков Лассо для задачи регрессии, а также метод отбора признаков Random Forests для задачи классификации.

Предлагается исследовать зависимость значения функции ошибки от размера доступной выборки, а также его дисперсию. В данной работе построена модель, аппроксимирующая значение функции ошибки с помощью заданного параметрического семейства функции. Оценка достаточного объёма выборки определяется по аппроксимации функции ошибки с использованием критерия, определяющего достаточный объём выборки по функции ошибки. Для вычислительного эксперимента предлагается использовать классические выборки из UCI репозитория и синтетические данные.

2. Постановка задачи

Дана выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x}_i \in \mathbb{R}^n$ — вектор признаков, $y_i \in \mathbb{Y}$.

Выборка является простой: элементы порождены независимо из одного распределения с фиксированными неизвестными параметрами, вероятность попадания каждого элемента в выборку одинакова. Предполагается, что выборка $\mathfrak{D}$ порождена согласно следующей гипотезе: модель, порождающая данные, задается в следующем виде:

$$y_i = f(\mathbf{x}_i, \mathbf{w}, \beta), \tag{1}$$

где $\mathbf{w} \in \mathbb{W} \subseteq \mathbb{R}^n$ — вектор параметров, β — дисперсия зависимой переменной. Зависимая переменная y аппроксимируется обобщенно линейной моделью:

$$\hat{y}_i = f(\mathbf{x}_i, \mathbf{w}) = \mu(\mathbf{w}^\top \mathbf{x}_i), \tag{2}$$

где μ — функция связи, для модели линейной регрессии:

$$\mu = \mathrm{id}, \tag{3}$$

для логистической регрессии:

$$\mu(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)}. \tag{4}$$

Предполагается, что при восстановлении параметров линейной регрессии (**??**), зависимая переменная порождается нормальным распределением:

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}(f(\mathbf{x}, \mathbf{w}), \hat{\beta}),$$

где $\hat{\beta}$ — выборочная дисперсия зависимой переменной y. Для модели логистической регрессии зависимая переменная порождается бернуллиевским распределением:

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{B}e\big(f(\mathbf{x}, \mathbf{w})\big).$$

Для модели f с вектором параметров $\mathbf{w}$ определим функцию правдоподобия и логарифмическую функцию правдоподобия выборки $\mathfrak{D}$:

$$L(\mathbf{w}|\mathfrak{D}) = \prod_{y,\mathbf{x}\in\mathfrak{D}} p(y|\mathbf{x}, \mathbf{w}), \quad l(\mathbf{w}|\mathfrak{D}) = \sum_{y,\mathbf{x}\in\mathfrak{D}} \log p(y|\mathbf{x}, \mathbf{w}). \tag{5}$$

Для множества индексов $\mathcal{A}$ независимой переменной $\mathbf{x} = [x_1, ..., x_n]^\top$, в векторы $\mathbf{x}_\mathcal{A}, \mathbf{w}_\mathcal{A}$ входят только те элементы, индексы которых принадлежат $\mathcal{A}$. Определим выборку $\mathfrak{D}_\mathcal{A}$ как множество независимых переменных $\mathbf{x}_\mathcal{A}$ и соответствующих им зависимых переменных y.

Для получения оптимального набора параметров и оценки вектора параметров используется принцип максимума правдоподобия:

$$\hat{\mathbf{w}}, \hat{\mathcal{A}} = \underset{\mathbf{w}\in\mathbb{W}, \mathcal{A}\subseteq\mathcal{J}}{\arg\max} L(\mathbf{w}_\mathcal{A}|\mathfrak{D}_\mathcal{A}), \tag{6}$$

где $\mathcal{J} = \{1, 2, ..., n\}$ — множество индексов.

В качестве функции ошибки используются следующие функции:

$$\begin{gathered} S_{\text{reg}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|}\sum_{\mathbf{x},y\in\mathfrak{D}} (y - f(\mathbf{x}, \mathbf{w}))^2, \\ S_{\text{class}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|}\sum_{\mathbf{x},y\in\mathfrak{D}} \big(y \ln f(\mathbf{x}, \mathbf{w}) + (1 - y)\ln(1 - f(\mathbf{x}, \mathbf{w}))\big). \end{gathered} \tag{7}$$

Определим функцию обратной экспоненты от функции ошибки (**??**):

$$e^{-S(\mathbf{w}|\mathfrak{D})}. \tag{8}$$

Заметим, что задача (**??**) эквивалентна задаче:

$$\hat{\mathbf{w}}, \hat{\mathcal{A}} = \underset{\mathbf{w}\in\mathbb{W}, \mathcal{A}\subseteq\mathbb{J}}{\arg\max}\, e^{-S(\mathbf{w}|\mathfrak{D}_\mathcal{A})}, \tag{9}$$

где S — функция ошибки (**??**).

Для нахождения оптимального набора признаков используется метод Лассо для задачи регрессии и метод Random Forests для задачи классификации. Требуется для каждого $n' \leq n$ найти оптимальный набор $\mathcal{A}_{n'}$ мощности n'. Используется тот факт, что с помощью данных методов все признаки упорядочиваются по значимости. Для того, чтобы выбрать набор признаков $\mathcal{A}_{n'}$ мощности $|\mathcal{A}_{n'}| = n'$ с наилучшей функцией ошибки, достаточно взять n' самых значимых признаков. Для получения оценки $\hat{\mathbf{w}}$ на наборе признаков $\mathcal{A}_{n'}$ решается оптимизационная задача (**??**).

Требуется по начально заданной выборке размера $m_0 \ll m$ получить оценку минимального объёма m^*, достаточного для выбора адекватной модели. Разделим выборку $\mathfrak{D}$ на обучающую и тестовую:

$$\mathfrak{D} = \mathfrak{D}_\mathcal{L} \sqcup \mathfrak{D}_\mathcal{T}. \tag{10}$$

Для введения понятия достаточности объёма выборки рассмотрим ожидаемое значение функции (??) по разным подвыборкам $\mathfrak{D}'_{\mathcal{L}}, \mathfrak{D}'_{\mathcal{T}}$ размера m' обучающей и тестовой выборок (??). Оценка вектора параметров $\hat{\mathbf{w}}$ является решением задачи (??) для выборки $\mathfrak{D}'_{\mathcal{L}}$:

$$l(m') = m'^{-1} \mathop{\mathsf{E}}_{\substack{\mathfrak{D}'_{\mathcal{L}} \subset \mathfrak{D}_{\mathcal{L}} \\ \mathfrak{D}'_{\mathcal{T}} \subset \mathfrak{D}_{\mathcal{T}} \\ |\mathfrak{D}'_{\mathcal{L}}| = |\mathfrak{D}'_{\mathcal{T}}| = m'}} e^{-S(\hat{\mathbf{w}}(\mathfrak{D}'_{\mathcal{L}})|\mathfrak{D}'_{\mathcal{T}})}. \tag{11}$$

Будем считать, что объем выборки m^* достаточен, если:

$$\forall m' > m^* \quad l(m') > (1-\delta) \max_{m > m^*} l(m),$$

где δ — достаточно малое пороговое значение.

2.1. Приближенное вычисление функции ошибки $\hat{l}(m)$

Для получения приближённого значения функции $l(m')$ введем процедуру бутстрепа:

1) Равновероятно генерируются случайные подвыборки $\mathfrak{D}'_{\mathcal{L}}, \mathfrak{D}'_{\mathcal{T}}$ размера m':
   - $\mathfrak{D}' \sim \mathcal{U}(\mathfrak{D})$ — вариант с полной информацией,
   - $\mathfrak{D}' \sim \mathcal{U}(\mathfrak{D}^0), \mathfrak{D}^0 \subset \mathfrak{D}, |\mathfrak{D}^0| = m'$ — вариант с неполной информацией.
2) Для полученых подвыборок вычисляется значение функции (??):
3) пп. 1-2 повторяются K раз, оценка $\hat{l}(m')$ функции $l(m')$ равняется среднему арифметическому среди всех полученных значений функции (??) на всех итерациях:

2.2. Аппроксимация функции ошибки $\phi(m) \sim \hat{l}(m)$

Для предсказания значения функции $\hat{l}(m)$ при $m > m_0$ введем параметрическое семейство функций:

$$\Phi = \{\phi(m) = a + b \cdot e^{c \cdot m} \mid a, b \in \mathbb{R}, c \in (-\infty, 0)\} \tag{12}$$

Аппроксимация функции $\hat{l}(m)$ является решением следующей задачи:

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \mathrm{MAE}(\hat{l}, \phi, 1, m_0), \tag{13}$$

где MAE — средняя абсолютная ошибка:

$$\mathrm{MAE}(\psi, \phi, m_1, m_2) = \frac{1}{m_2 - m_1 + 1} \sum_{i=m_1}^{m_2} |\phi(i) - \psi(i)|. \tag{14}$$

3. Вычислительный эксперимент

3.1. Эксперимент на синтетических выборках

Пусть $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$ — набор векторов-столбцов. Построим зависимость (**??**) для различных конфигураций выборок:

1) Случайная выборка ($m = 200, n = 10$):

$$\boldsymbol{\chi}_i \sim \mathcal{U}(\mathbf{5}, \mathbf{6}),\ \mathbf{w} \sim \mathcal{U}(\mathbf{0}, \mathbf{1}),\ y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, 1)$.

2) Скоррелированная выборка ($m = 200, n = 10$):

$$\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_3 \sim \mathcal{N}(\mathbf{1}, \mathbf{1}),\ \boldsymbol{\chi}_4, ..., \boldsymbol{\chi}_6 \sim \boldsymbol{\chi}_0 + \varepsilon,\ \boldsymbol{\chi}_7, ..., \boldsymbol{\chi}_{10} \sim \boldsymbol{\chi}_1 + \varepsilon,$$

$$y = 0.3x_1 + 0.7x_2 + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, 1)$,

3) Ортогональная выборка ($m = 200, n = 10$):

$$\{\boldsymbol{\chi_i} \in \mathbb{R}^m\}_{i=1}^n \text{ — ортогональный набор},\ \mathbf{w} \sim U(\mathbf{1}, \mathbf{1}),\ y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, 0.5)$.

4) Избыточная выборка ($m = 200, n = 10$):

$$\boldsymbol{\chi}_i \sim \mathcal{N}(1, 1),$$

$$w_i \sim \mathcal{U}(1, 1),\ i \leq 5,$$

$$w_i = 0,\ i > 5,$$

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, 0.5)$

На рис. 1 представлена функция (**??**), посчитанная с помощью бутстрепа в варианте с полной информацией для различного числа признаков n'. Дисперсия функции (**??**) монотонно уменьшается с увеличением размера обучающей

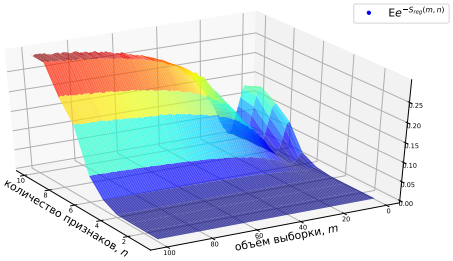
выборки. Поведение функции (**??**) при различных n' соответствует ожиданиям. Видно, что для всех выборок функция (**??**) хорошо аппроксимируется семейством функций (**??**).

В табл. 1 приведены оптимальные значения m^*, n^* для разных конфигураций выборок. Полученнные значения n^* ожидаемы из способа генерации. Значения m^* предстоит предсказать по построенной аппроксимации функции $\hat{l}(m)$ при заданном m_0.

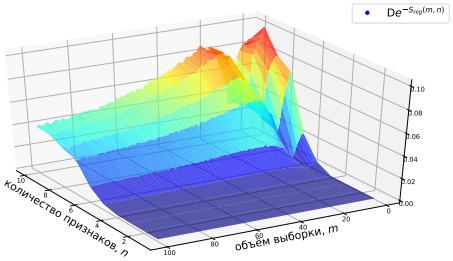
Таблица 1: Оптимальные значения m^*, n^* для различных конфигураций выборок

Выборка	m^*	m	n^*	n	$\mathsf{D}\varepsilon$
Случайная выборка	72	100	10	10	1
Скоррелированная выборка	31	100	2	10	1
Ортогональная выборка	45	100	10	10	0.5
Избыточная выборка	22	100	5	10	0.5

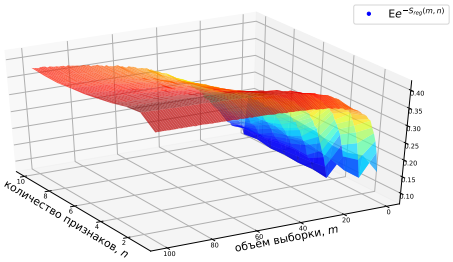
Матожидание **Дисперсия**



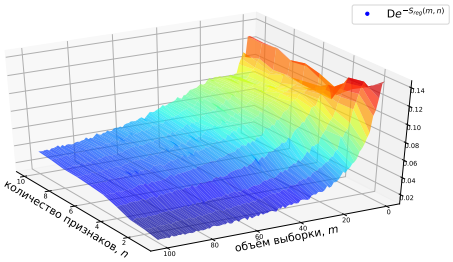
(a) Случайная выборка



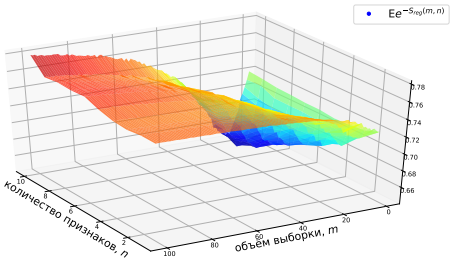
(b) Случайная выборка



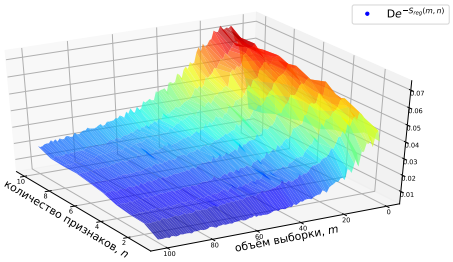
(c) Скоррелированная выборка



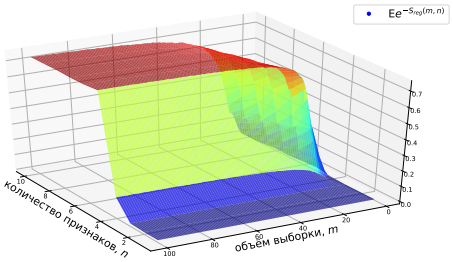
(d) Скоррелированная выборка



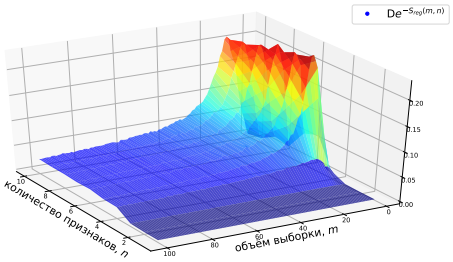
(e) Ортогональная выборка



(f) Ортогональная выборка



(g) Избыточная выборка

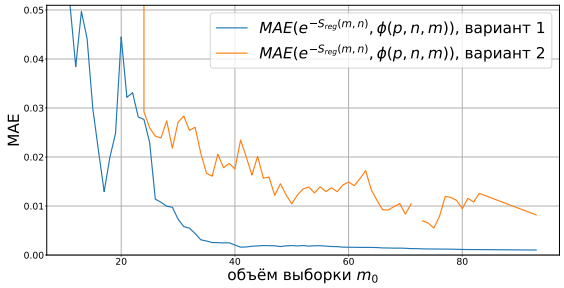


(h) Избыточная выборка

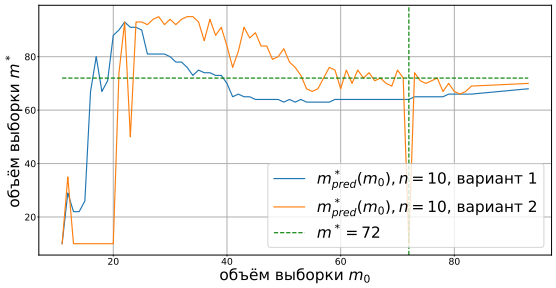
Рис. 1: Зависимость значения функции $e^{-S(m,n)}$ от объема выборки m и количества параметров n для синтетических выборок

На рис. 2. представлены графики качества аппроксимации функции (**??**), а также предсказание $\hat{m^*}$ при различных m_0 для разных конфигураций выборок. Для аппроксимации функции $\hat{l}(m)$ вариант с неполной информацией дает большую ошибку, чем вариант с полной информацией, и это ожидаемый результат: чем меньше информации, тем хуже качество предсказания. При $m_0 \to m$ аппроксимация в варианте с неполной информацией стремится к аппроксимации в варианте с полной информацией, поэтому при больших m_0 предсказания $\hat{m^*}$ для этих двух вариантов практически совпадают. Для случайной и избыточной выборок получилось построить адекватное предсказание при $m_0 < m^*$.

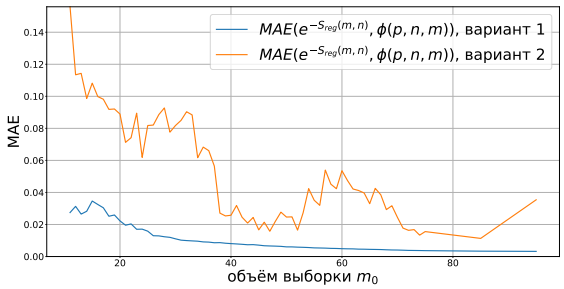
Аппроксимация $e^{-S(m,n)}$ **Предсказание m^***



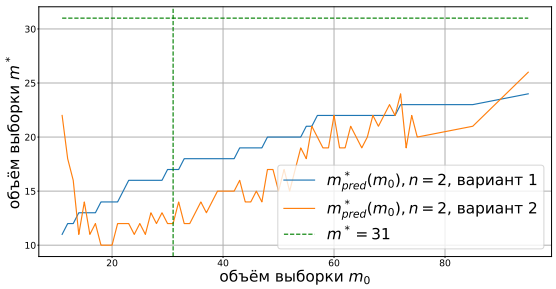
(a) Случайная выборка



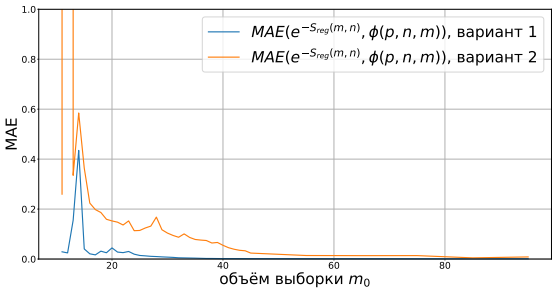
(b) Случайная выборка



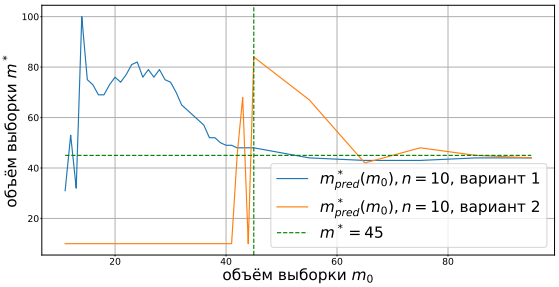
(c) Скоррелированная выборка



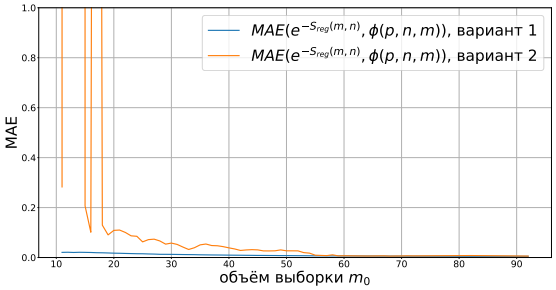
(d) Скоррелированная выборка



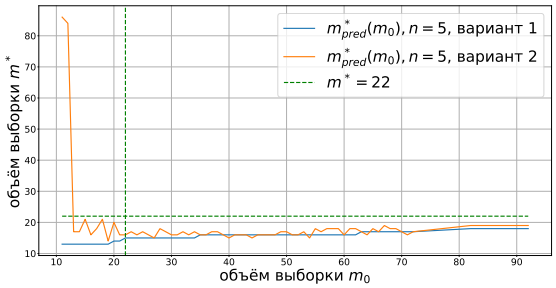
(e) Ортогональная выборка



(f) Ортогональная выборка



(g) Избыточная выборка



(h) Избыточная выборка

Рис. 2: Качество предсказания $e^{-S(m,n)}$ и m^* в зависимости от объема обучающей выборки m_0 для синтетических выборок

3.2. Эксперимент на выборках из UCI репозитория

Для эксперимента использовались выборки из UCI репозитория, описанные в табл. 2.

Таблица 2: Выборки из UCI репозитория

Выборка	Тип задачи	m	n
Diabetes	регрессия	442	11
Boston	регрессия	506	14
Wine	классификация	130	14
Nba	классификация	400	20

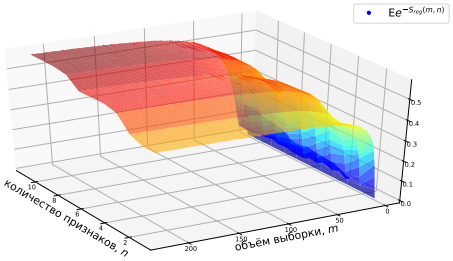
На рис. 3 представлена функция $\hat{l}(m)$, посчитанная с помощью бутстрепа в варианте с полной информацией для различного числа признаков n'. Качественное поведение функции $\hat{l}(m)$ похоже на поведение данной функции для синтетических выборок и также хорошо аппроксимируется семейством функций (**??**).

В табл. 3 приведены оптимальные значения m^*, n^* для различных выборок. Примечательно, что для задачи классификации достаточный объём выборки меньше, чем для задачи регрессии.

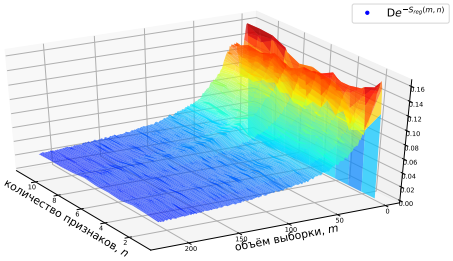
Таблица 3: Оптимальные значения m^*, n^* для различных синтетических выборок

Выборка	m^*	m	n^*	n
Diabetes	96	221	11	11
Boston	102	253	14	14
Wine	27	65	14	14
Nba	38	200	2	20

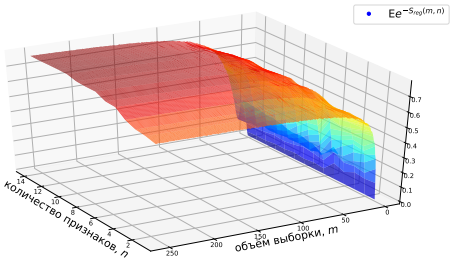
Матожидание **Дисперсия**



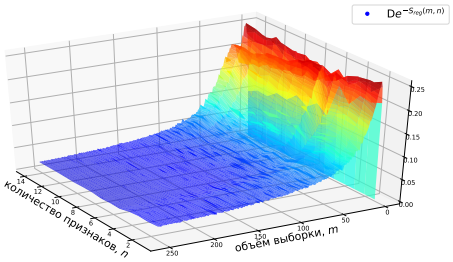
(a) Diabetes



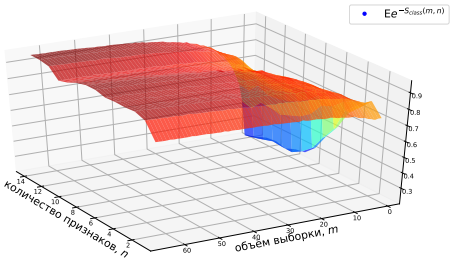
(b) Diabetes



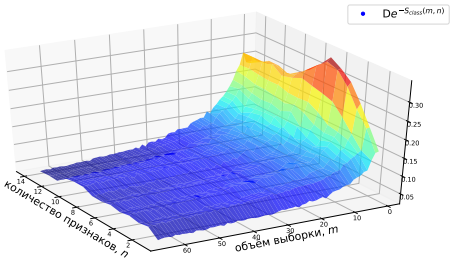
(c) Boston



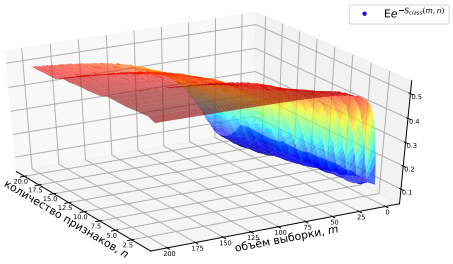
(d) Boston



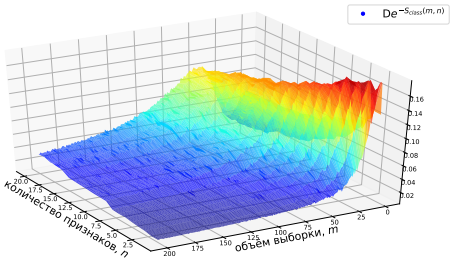
(e) Wine



(f) Wine



(g) Nba

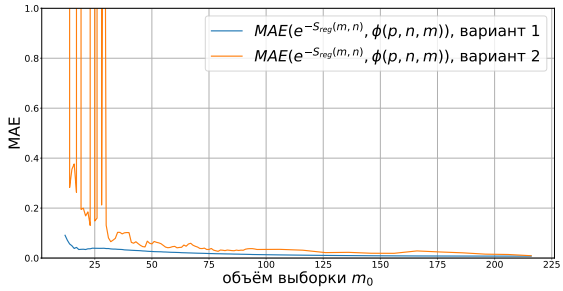


(h) Nba

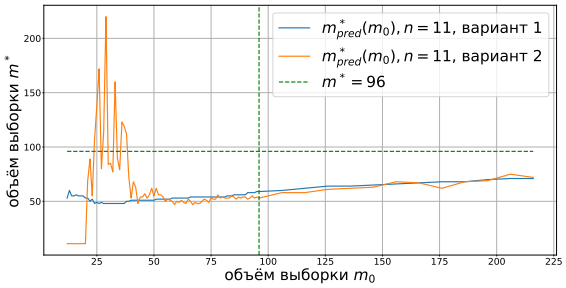
Рис. 3: Зависимость значения функции $e^{-S(m,n)}$ от объема выборки m и количества параметров n для выборок из UCI репозитория

На рис. 4. представлены графики качества аппроксимации функции $\hat{l}(m)$, а также предсказание $\hat{m^*}$ при различных m_0 для разных выборок. Так же, как и для синтетических выборок, аппроксимация функции $\hat{l}(m)$ в варианте с неполной информацией даёт качество хуже, чем в варианте с полной информацией. Также качество аппроксимации в варианте с неполной информацией сильно шумит при малых значениях m_0. При $m_0 \to m$ аппроксимация в варианте с неполной информацией стремится к аппроксимации в варианте с полной информацией, поэтому при больших m_0 предсказания $\hat{m^*}$ для этих двух вариантов практически совпадают. Для выборок задачи регрессии получилось построить адекватное предсказание достаточного объёма выборки $\hat{m^*}$ при $m_0 < m^*$. Для выборок задачи классификации при $m_0 < m^*$ построить адекватное предсказание $\hat{m^*}$ не получилось. Это может быть связано с тем, что достаточный объём m^* для задачи классификации сильно меньше, чем для задачи регрессии, а при меньших m_0 аппроксимация функции $\hat{l}(m)$ хуже, чем при больших m_0.

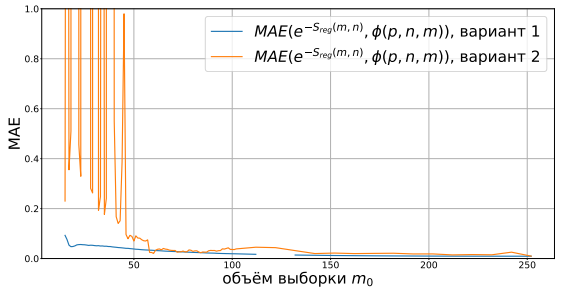
Аппроксимация $e^{-S(m,n)}$ **Предсказание m^***



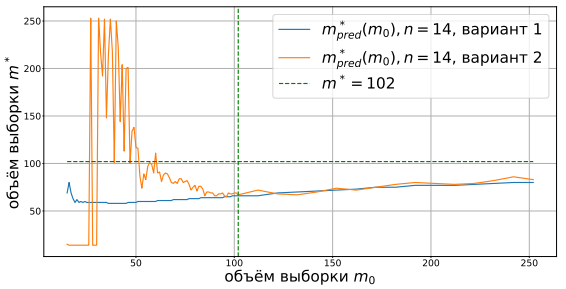
(a) Diabetees



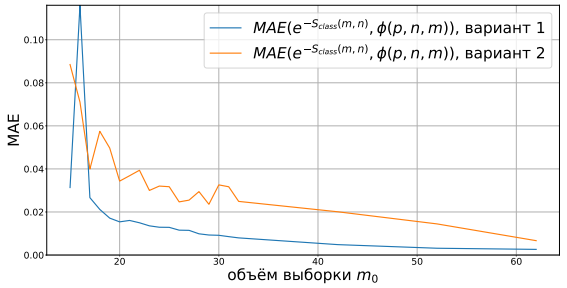
(b) Diabetees



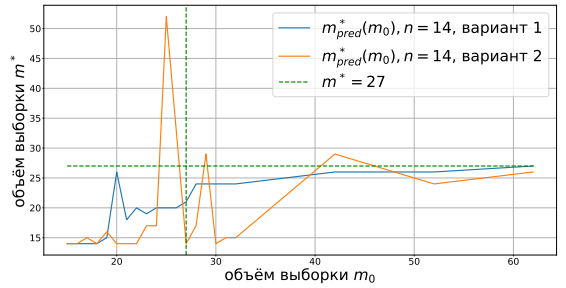
(c) Boston



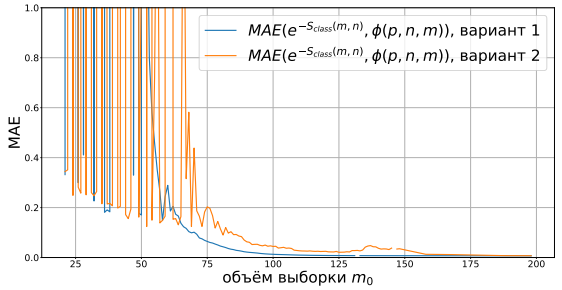
(d) Boston



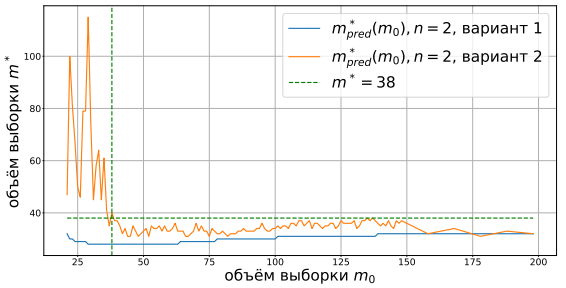
(e) Wine



(f) Wine



(g) Nba



(h) Nba

Рис. 4: Качество предсказания $e^{-S(m,n)}$ и m^* в зависимости от объема обучающей выборки m_0 для выборок из UCI репозитория

4. Заключение

Построен алгоритм раннего прогнозирования достаточного объёма выборки для обобщённой линейной модели с неизвестным оптимальным набором признаков. Реализован один из методов выбора оптимального набора признаков заданного размера. Для некоторых выборок задача раннего прогнозирования была успешно решена. Данный алгоритм можно улучшить с помощью использования в построении модели аппроксимации $\phi(m) \sim \hat{l}(m)$ свойств оценки вектора параметров $\hat{\mathbf{w}}$ обобщенной линейной модели.

Список литературы

[1] S. G. Self and R. H. Mauritsen Power/sample size calculations for generalized linear models // Biometrics, 1988. Vol. 44. P. 79-86.

[2] G. Shieh On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000. Vol. 56. P. 1192-1196.

[3] G. Shieh On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005. Vol. 128. P. 43-59.

[4] Fei Wang and Alan E. Gelfand A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models // Statistical Science, 2002. Vol. 17. P. 193-208.

[5] V. V. Strijov Error function in regression analysis // Factory Laboratory, 2013.

[6] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula and Long H Ngo Predicting sample size required for classification performance // BMC Medical Informatics and Decision Making.

[7] Kevin K. Dobbin, Yingdong Zhao, and Richard M. Simon How Large a Training Set is Needed to Develop a Classifier for Microarray Data? // American Association for Cancer Research, 2008.

[8] R. Tibshirani Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society, 1996. Vol. 32. P. 267-288.

[9] T. Hastie, J. Taylor, R. Tibshirani, G. Walter Forward stagewise regression and the monotone lasso // Electronic journal of Statistics, 2007. Vol. 1. P 1-29.

[10] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression // The Annals of Statistics, 2004. Vol. 32. P. 407-499.

[11] Leo Brieman Random forests // Machine Learning, 2001. P. 5-32.