Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели

Валентин Бучнев

Московский физико-технический институт
Физтех-школа прикладной математики и информатики

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2020 г.

Задача раннего прогнозирования достаточного объёма выборки

Цель исследования

Предложить метод предсказания достаточного объема выборки для обобщенной линейной модели на ранних этапах сбора данных.

Проблема

Большинство методов требуют заведомо избыточного объема выборки. Неизвестна структура модели — состав признаков, известен лишь класс модели.

Метод решения

Оценка объема строится по собранной выборке путем анализа свойств аппроксимации параметрическим семейством функций эмперической функции ошибки обобщенной линейной модели.

Существующие методы

Ассимптотические методы

- S. G. Self and R. H., Mauritsen Power/sample size calculations for generalized linear models // Biometrics, 1988
- G. Shieh, On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000.
- G. Shieh On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005.

Байесовские методы

- D. B. Rubin and H. S. Stern Sample size determination using posterior predictive distributions // Sankhya : The Indian Journal of Statistics Special Issue on Bayesian Analysis, 1998.

Существующие методы оценки объёма выборки

Подход	Формула
Метод доверительных интервалов $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : EX = \mu$	$m = \left(\frac{z_{\alpha/2}\sigma}{\bar{X}-\mu}\right)^2$
Тест на равенство: $Z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : p = p_0$ против $H_1 : p \neq p_0$	$m = \frac{(z_{\mathbf{Pow}}+z_{\alpha/2})^2 p(1-p)}{(\hat{p}-p_0)^2}$
Тест отношения правдоподобия: $\gamma_m : \chi^2_{p,1-\mathrm{Pow}}(\gamma_m) = \chi^2_{p,\alpha}$	$m = \frac{\gamma_m}{\Delta^*}$, где $\Delta^* = \mathrm{E}_X\left[\frac{-X(\beta-\beta^*)}{1+e^{-X\beta}} - \log\left(\frac{1+e^{-X\beta}}{1+e^{-X\beta^*}}\right)\right]$
Статистика Вальда: $Z = \frac{\hat{\beta}-\beta^0}{\sqrt{\hat{V}}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : \beta = \beta^0$	$\hat{m} = \frac{\left(\sqrt{V_1}z_{\mathbf{Pow}} - \sqrt{V_0}z_{\alpha/2}\right)^2}{(\beta^1-\beta^0)^2}$
Заданная точность регрессии: $\hat{\beta}_j = t_{1-\alpha/2}(m-n-1)\sqrt{\frac{1-\mathbf{R}^2}{(1-\mathbf{R}_j^2)(m-n-1)}}$	$m^* = \frac{z^2_{\alpha/2}}{\delta^2}\left(\frac{1-\mathbf{R}^2}{1-\mathbf{R}_j^2}\right)\left(\frac{\chi^2_{1-\gamma}(m-1)}{m-n-1}\right) + n +$ 1, где $\mathbf{R} = \boldsymbol{\rho}'_{yx}\mathbf{R}^{-1}_{xx}\boldsymbol{\rho}_{yx}$
С помощью метода Bootstrap	$m = \left(\frac{z_{\alpha/2}\sigma}{\bar{X}-\mu}\right)^2$ и $m = \frac{z^2_{\alpha/2}}{(\bar{X}-\mu)^2}\left(\frac{1-R^2}{1-R_j^2}\right) + n$

Постановка задачи раннего прогнозирования

Дано

Выборка размера m: $\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^m$,
где $\mathbf{x}_i \in \mathbb{R}^n$ - вектор признаков, $y_i \in \mathbb{Y}$.

Обобщённая линейная модель

Зависимая переменная y аппроксимируется обобщенной линейной моделью:

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

$$\mu = \mathrm{id} \text{ — задача регрессии,}$$

$$\mu(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)} \text{ — задача классификации}$$

Постановка задачи раннего прогнозирования

Функция ошибки $S(\mathbf{w}, \mathfrak{D})$ для задач регрессии и классификации

$$S_{\text{reg}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|} \sum_{\mathbf{x},y \in \mathfrak{D}} (y - f(\mathbf{x}, \mathbf{w}))^2,$$

$$S_{\text{class}}(\mathbf{w}|\mathfrak{D}) = \frac{1}{|\mathfrak{D}|} \sum_{\mathbf{x},y \in \mathfrak{D}} \big(y \ln f(\mathbf{x}, \mathbf{w}) + (1 - y) \ln(1 - f(\mathbf{x}, \mathbf{w}))\big).$$

Функция ошибки

Будем рассматривать ожидаемое значение функции $e^{-S(\hat{\mathbf{w}}(\mathfrak{D}_\mathcal{L})|\mathfrak{D}_\mathcal{T}))}$ по разным обучающим и тестовым выборкам размера m:

$$I(m) = \mathsf{E} e^{-S(\hat{\mathbf{w}}(\mathfrak{D}_\mathcal{L})|\mathfrak{D}_\mathcal{T}))}.$$

Постановка задачи раннего прогнозирования

Функция правдоподобия

Определим функцию правдоподобия и логарифмическую функцию правдоподобия выборки $\mathfrak{D}$:

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

где $p(y|\mathbf{x}, \mathbf{w})$ — плотность зависимой переменной.

Оценка вектора параметров и оптимальный набор признаков

Для получения оптимального набора признаков и оценки $\hat{\mathbf{w}}$ используется принцип максимума правдоподобия:

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

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

Постановка задачи раннего прогнозирования

Функция ошибки $l(m)$

$\hat{l}(m)$ — оценка функции $l(m)$, посчитанная с помощью метода бутстреп по разным обучающим и тестовым подвыборкам размера m выборки $\mathfrak{D}$.

Бутстреп

Есть 2 варианта генерации подвыборок:

- $\mathfrak{D}' \sim \mathcal{U}(\mathfrak{D})$ — вариант 1, с полной информацией,
- $\mathfrak{D}' \sim \mathcal{U}(\mathfrak{D}^0), \mathfrak{D}^0 \subset \mathfrak{D}, |\mathfrak{D}^0| = m'$ — вариант 2, с неполной информацией.

Предлагаемый метод решения

Семейство функций Φ

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

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

Аппроксимация $\phi(m) \sim \hat{I}(m)$

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

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

где

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

Критерий достаточности объёма

Критерий достаточности объема

Определим достаточный объем выборки m^* как наименьший объём, что

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

где δ — достаточно малое пороговое значение.

Оценка $\hat{m}^*$

$\hat{m}^*$ — наименьший объём выбрки, что

$$\hat{\phi}(\hat{m}^*) > (1-\delta) \max_{m' > \hat{m}^*} \hat{\phi}(m').$$

Вычислительный эксперимент

Цель эксперимента

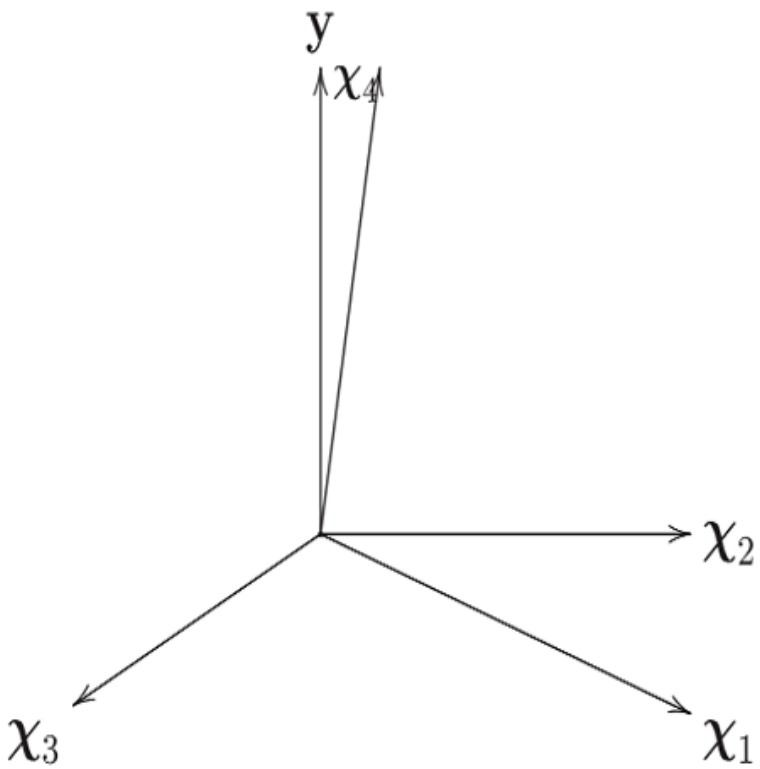
Проверить работоспособность предложенного метода.

План эксперимента

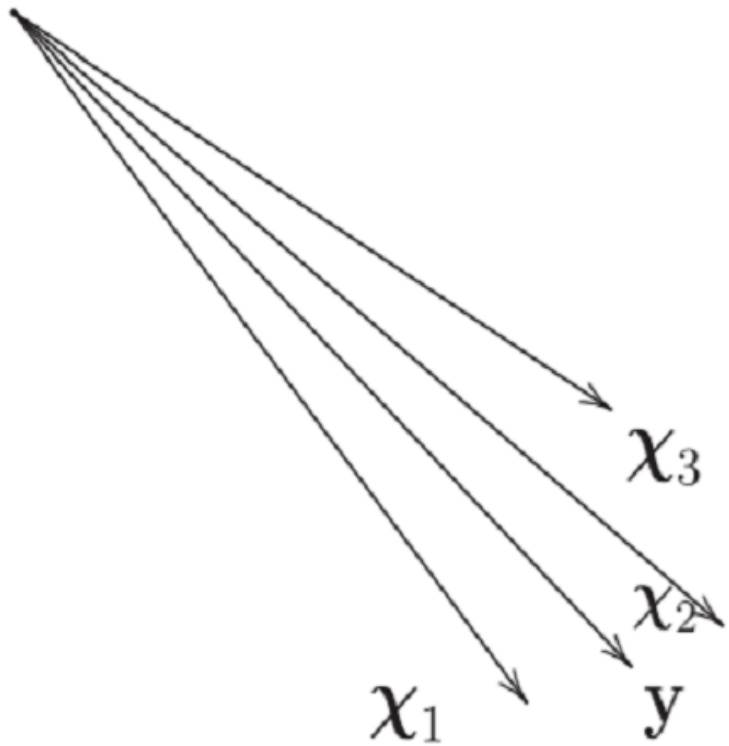
1. Вычисление значимостей признаков, определение оптимальных наборов $\mathcal{A}(n')$ размера n'.
2. Приближенное вычисление функции ошибки $\hat{l}(m)$ с помощью метода бутстреп, получение значения достаточного объёма выборки m^*.
3. Построение модели аппроксимации функции ошибки $\hat{l}(m)$ параметрическим семейством функций Ф, получение аппроксимации функции ошибки $\hat{\phi}(m)$.
4. Получение оценки достаточного объёма выборки $\hat{m}^*$.

Синтетические выборки

Пусть $\mathbf{X} = [\chi_1, ..., \chi_n]$ — набор векторов-столбцов.



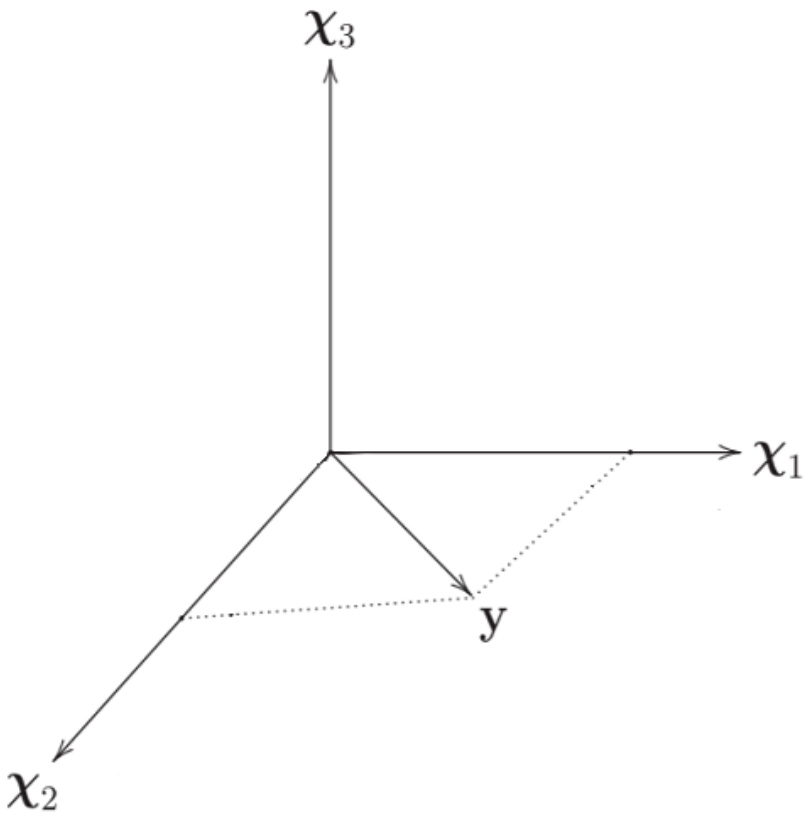
(a) Случайная выборка



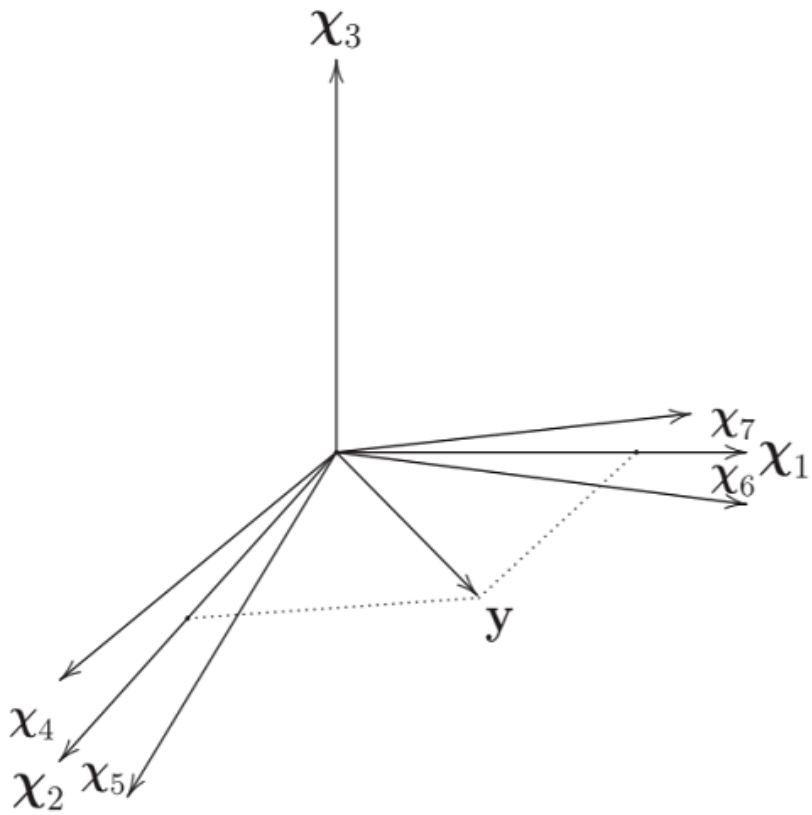
(b) Скоррелированная выборка

Синтетические выборки

Пусть $\mathbf{X} = [\chi_1, ..., \chi_n]$ — набор векторов-столбцов.



(a) Ортогональная выборка



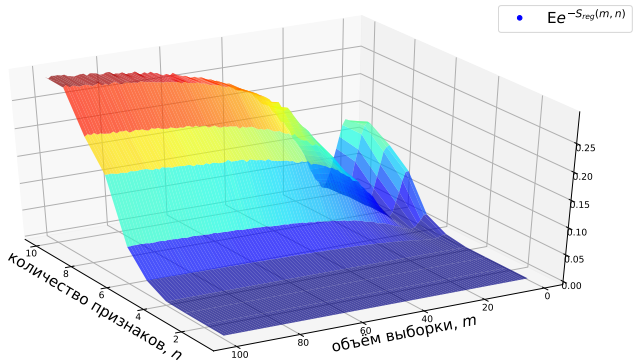
(b) Избыточная выборка

Синтетические выборки

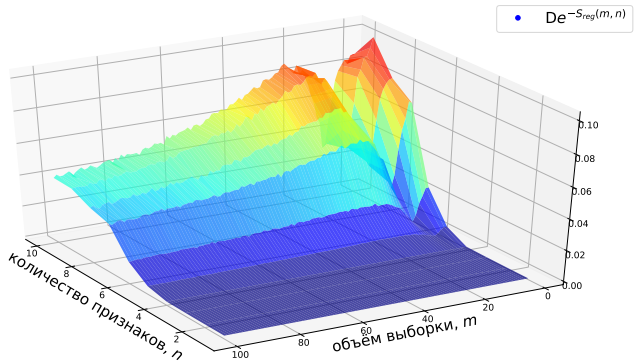
Выборка	m^*	m	n^*	n	Dε
Случайная выборка	72	100	10	10	1
Скоррелированная выборка	31	100	2	10	1
Ортогональная выборка	45	100	10	10	0.5
Избыточная выборка	22	100	5	10	0.5

Результаты эксперимента на синтетических выборках

Матожидание



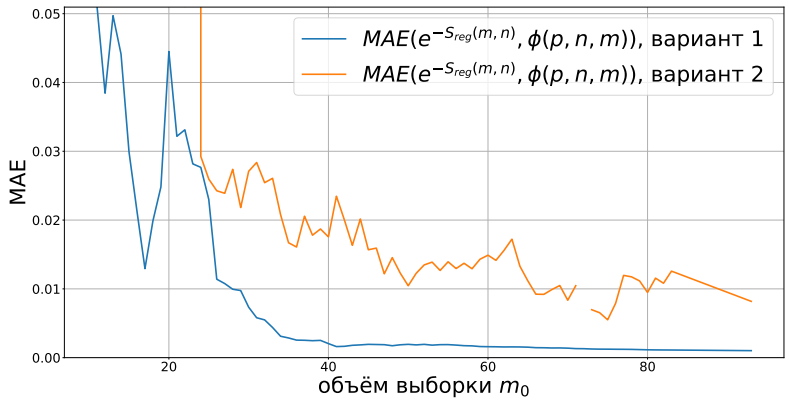
Дисперсия



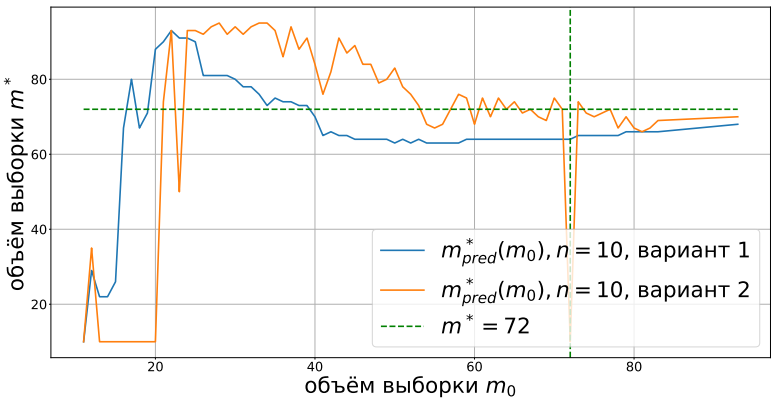
Зависимость значения функции $\hat{I}(m, n)$ от объема выборки m и количества параметров n для случайной выборки

Результаты эксперимента на синтетических выборках

Аппроксимация $e^{-S(m,n)}$



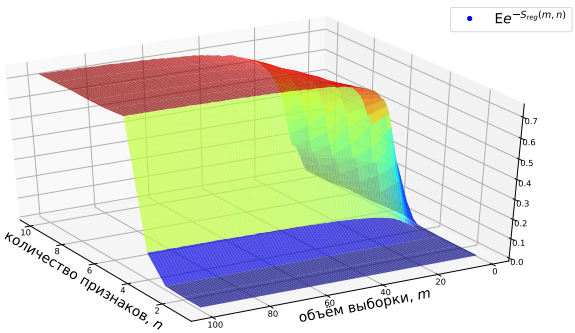
Предсказание $\hat{m}^*$



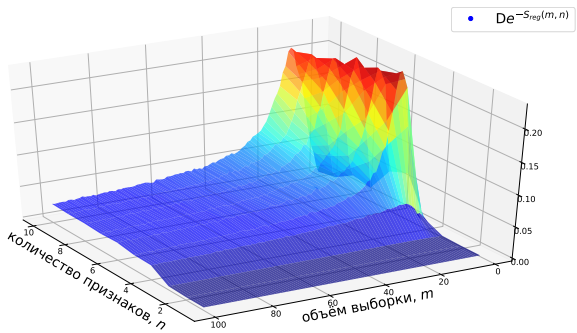
Качество предсказания $\hat{I}(m,n)$ и m^* в зависимости от объема обучающей выборки m_0 для случайной конфигурации выборки

Результаты эксперимента на синтетических выборках

Матожидание



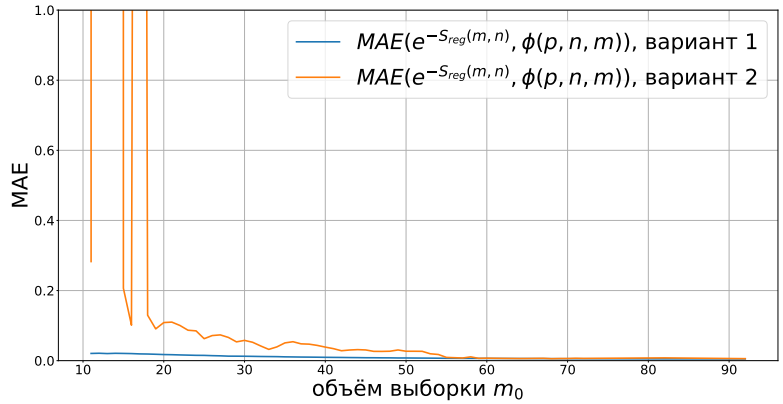
Дисперсия



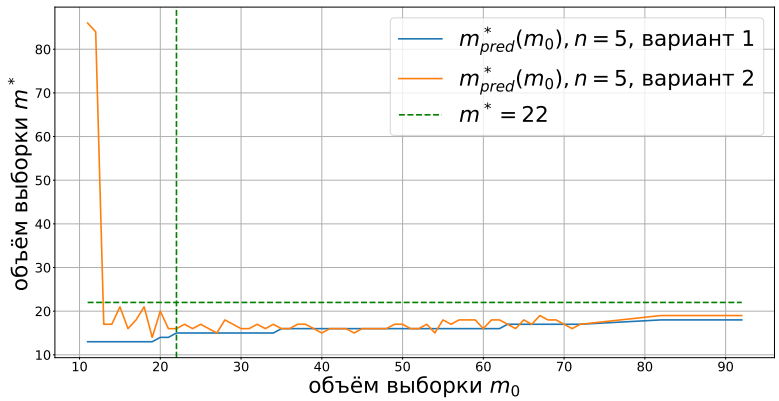
Зависимость значения функции $\hat{I}(m, n)$ от объема выборки m и количества параметров n для случайной выборки

Результаты эксперимента на синтетических выборках

Аппроксимация $e^{-S(m,n)}$



Предсказание $\hat{m}^*$



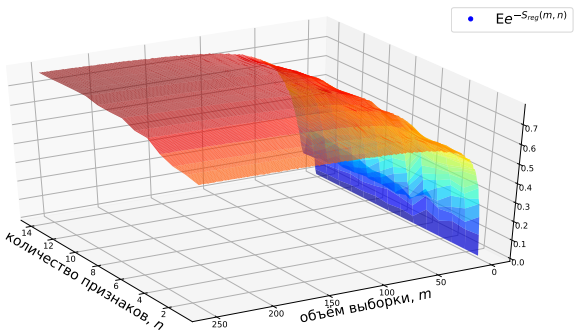
Качество предсказания $\hat{I}(m, n)$ и m^* в зависимости от объема обучающей выборки m_0 для избыточной конфигурации выборки

Выборки из UCI репозитория

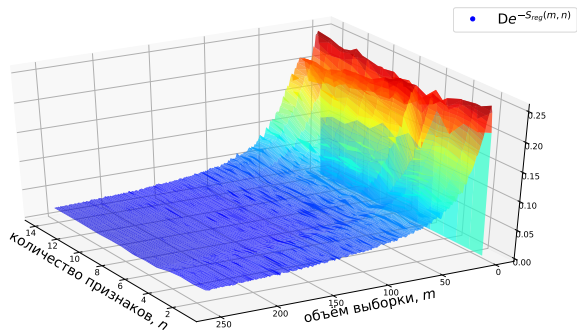
Выборка	m^*	m	n^*	n
Diabetes	96	221	11	11
Boston	102	253	14	14
Wine	27	65	14	14
Nba	38	200	20	20

Результаты эксперимента на выборках из UCI репозитория

Матожидание



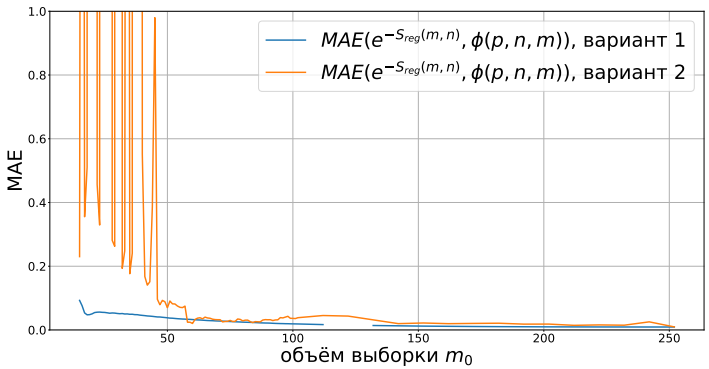
Дисперсия



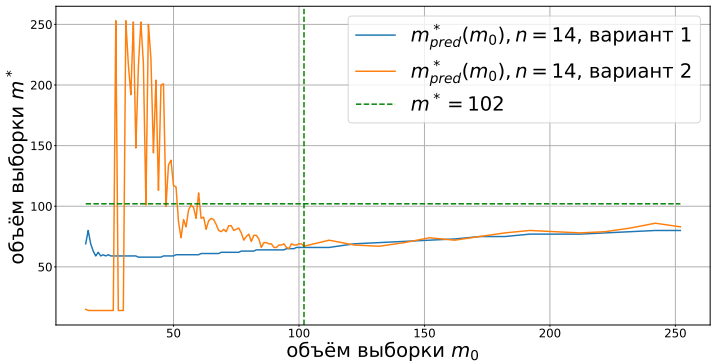
Зависимость значения функции $\hat{I}(m, n)$ от объема выборки m и количества параметров n для выборки Boston

Результаты эксперимента на выборках из UCI репозитория

Аппроксимация $e^{-S(m,n)}$



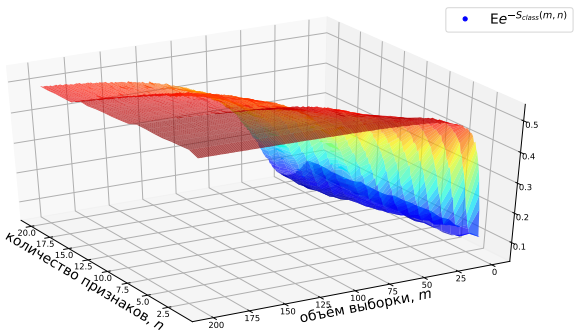
Предсказание m^*



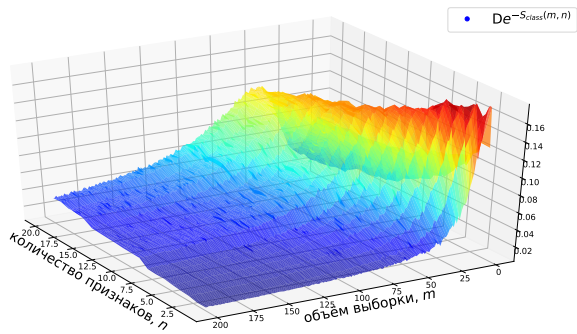
Качество предсказания $\hat{I}(m, n)$ и m^* в зависимости от объема обучающей выборки m_0 для выборки Boston

Результаты эксперимента на выборках из UCI репозитория

Матожидание



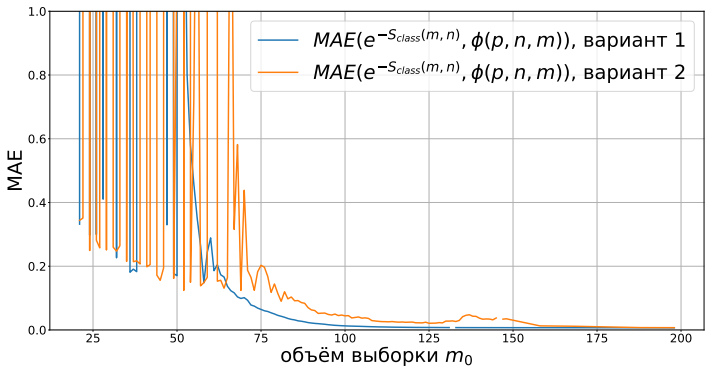
Дисперсия



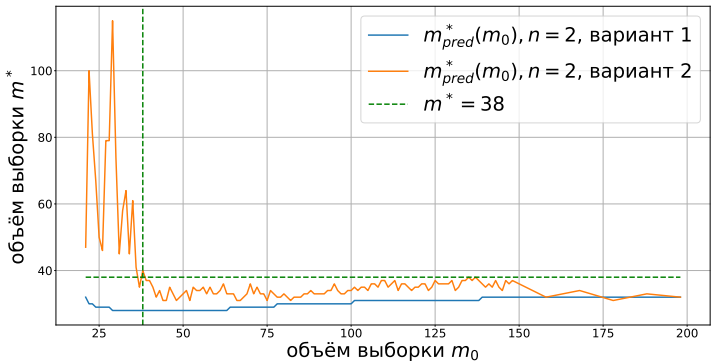
Зависимость значения функции $\hat{I}(m, n)$ от объема выборки m и количества параметров n для выборки Nba

Результаты эксперимента на выборках из UCI репозитория

Аппроксимация $e^{-S(m,n)}$



Предсказание m^*



Качество предсказания $\hat{I}(m, n)$ и m^* в зависимости от объема обучающей выборки m_0 для выборки Nba

Заключение

- Задача прогнозирования достаточного объема выборки сведена к задаче аппроксимации функции ошибок.
- Показана работоспособность предложенного метода на синтетических выборках, а также на выборках из UCI репозитория.