Численные методы оценки оптимального объёма выборки для логистической и линейной регрессии

Гадаев Тамаз

Московский Физико-Технический Институт
Физтех-школа прикладной математики и информатики
Кафедра Интеллектуальные системы
Научный руководитель: д. ф.-м. н. Стрижов В. В.

Математические методы распознавания образов - 2019

28 ноября 2019 г.

Цель исследования

Проблема

Завышенные оценки оптимального объёма выборки приводят к потерям ресурсов в случаях, когда сбор выборки затратен

Цель

Точно оценить оптимальный объём выборки для построения обобщённо-линейной модели

Предлагается

Провести исследование методов на основе статистического анализа мощности, байесовского вывода и эвристик

Литература

Методы основанные на анализе мощности статистических критериев

- S. G. Self and R. H. Mauritsen Power/sample size calculations for generalized linear models // Biometrics, 1988. Vol. 44. P. 79–86.
- G. Shieh On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000. Vol. 56. P. 1192–1196

Методы основанные на Байесовском выводе

- L. Joseph and R. du Berger and P. Be'lisle Bayesian and mixed bayesian likelihood criteria for sample size determination // Statistician, 1995. Vol. 16. P. 769–781.
- A. Motrenko and V. Strijov and W. Weber Sample size determination for logistic regression // Journal of Computational and Applied Mathematics, 2014. Vol. 255. P. 743–752.
- 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. Vol. 60. P. 161–175.

Эвристичекие методы

- Maher Qumsiyeh Using the bootstrap for estimation the sample size in statisticalexperiments // Journal of modern applied statistical methods, 2002. Vol. 8(3). P. 305P321.

Дана выборка

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

по $\mathfrak{D}_m$ восстанавливается зависимость $f(\boldsymbol{x}) = y$, где

$$f(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}\boldsymbol{w}^T, & \text{если } y \in \mathbb{R}, \quad \text{(линейная регрессия)} \\ \mathbb{I}[\sigma(\boldsymbol{x}\boldsymbol{w}^T) > 0.5], & \text{если } y \in \{0,1\}, \quad \text{(логистическая регрессия)} \end{cases}$$

Оценка вектора параметров $\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \ln L(\mathfrak{D}, \boldsymbol{w})$, где

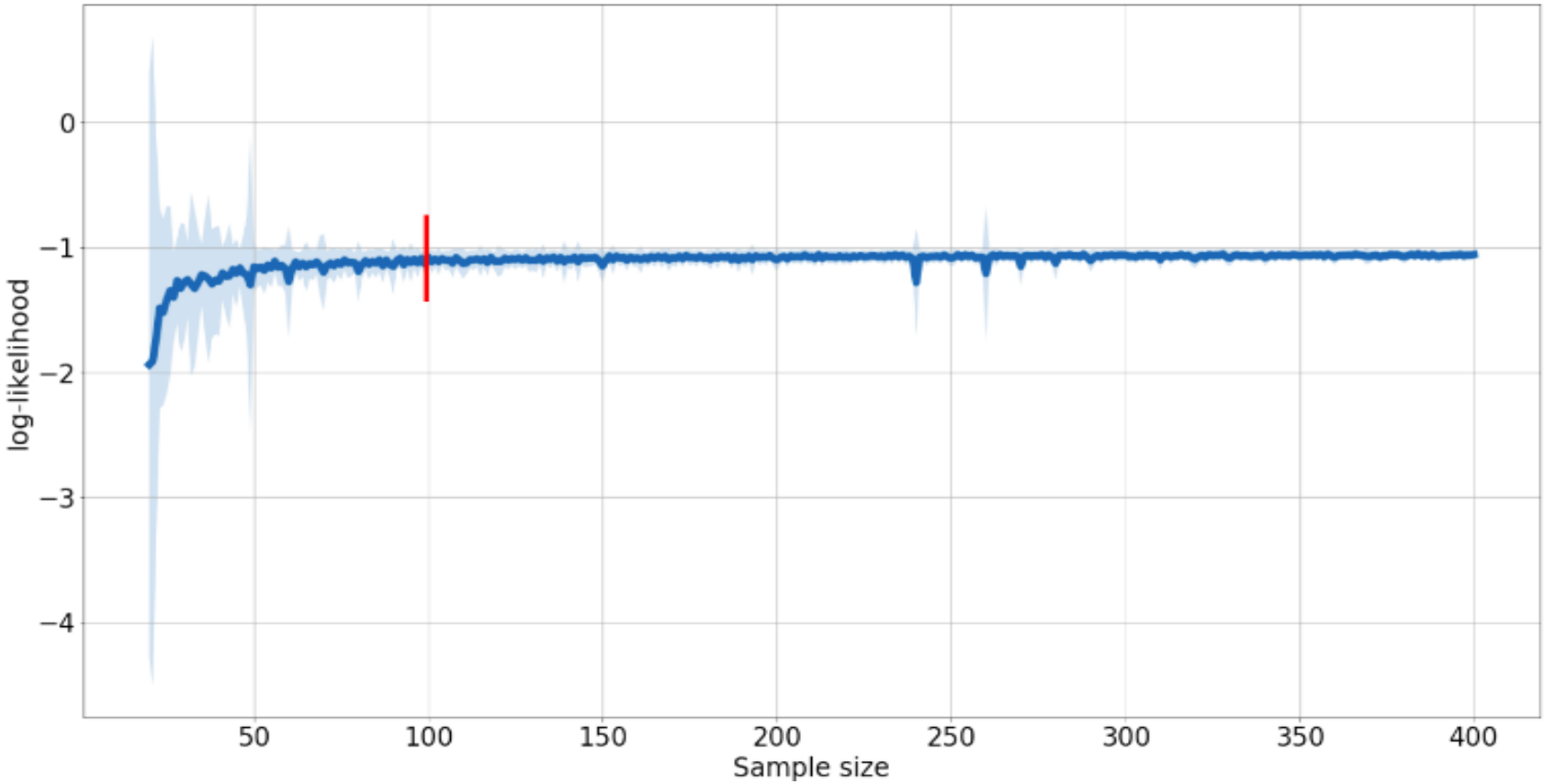
$$lnL(\mathfrak{D}, \boldsymbol{w}) = \begin{cases} \sum_{i=1}^m (y_i - f(\boldsymbol{x}_i))^2, & \text{если } y \in \mathbb{R} \\ \sum_{i=1}^m y_i ln(f(\boldsymbol{x}_i)) + (1 - y_i) ln(1 - f(\boldsymbol{x}_i)), & \text{если } y \in \{0,1\} \end{cases}$$

Задача

Необходимо адекватно определить оптимальность объёма выборки для нахождения $\hat{\boldsymbol{w}}$ и в соответствии с этим определением сделать точную оценку m^*

Пример: оценка по дисперсии логарифма правдоподобия

В этом примере оптимальным выбирается такой размер выборки, при котором дисперсия логарифма правдоподобия выходит на константу.



Модель порождения данных:

$$p(y|\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}_u, \boldsymbol{w}_v) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c\,(y, \phi)\right)$$

Вектор признаков $\boldsymbol{x}$ представлен как $\boldsymbol{x} = [\boldsymbol{u}, \boldsymbol{v}]$, вектор параметров $\boldsymbol{w} = [\boldsymbol{w}_u, \boldsymbol{w}_v]$, истинный вектор параметров $\boldsymbol{m} = [\boldsymbol{m}_u, \boldsymbol{m}_v]$.
Рассмотрим гипотезу:

$$H_0 : \boldsymbol{m}_u = \boldsymbol{m}_u^0, \quad H_1 : \boldsymbol{m}_u \neq \boldsymbol{m}_u^0$$

где $\boldsymbol{m}_u^0$ — гиперпараметр метода.

Основная идея

В каждом из статистических методов вычисляется статистика $T(\mathfrak{D}_m, \boldsymbol{w}_u^0)$. Известно, что $T \sim \chi_k^2$ при H_0 и $T \sim \chi_k^2(\gamma)$ при H_1, где γ в каждом методе вычсиляется по-своему.
Условие оптимальности:
$[P(H_0 \text{ отвергнута}|H_0) = \alpha] \cap [P(H_0 \text{ отвергнута}|H_1) = \beta]$
Найти m^* можно решив уравнение

$$\chi^2_{k,1-\alpha} = \chi^2_{k,\beta}\left(\gamma(m)\right)$$

Пусть статистики $S_{m,u}(\mathbf{w}_u, \mathbf{w}_v)$ и $S_{m,v}(\mathbf{w}_u, \mathbf{w}_v)$ - производные логарифма правдоподобия выборки $\mathfrak{D}_m$ по параметрам $\mathbf{w}_u$ и $\mathbf{w}_v$ соответственно.
Рассмотрим $\mathbf{s}_m = S_{m,u}\left(\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0\right)$, где $\hat{\mathbf{w}}_v^0$ определяется из уравнения

$$S_{m,v}\left(\mathbf{m}_u^0, \mathbf{w}_v\right) = 0$$

Метод множителей Лагранжа использует статистику $LM = \mathbf{s}_m^{\mathsf{T}} \mathbf{Q}_m^{-1} \mathbf{s}_m$, где $\mathbf{Q}_m$ — матрица ковариации $\mathbf{s}_m$.
$LM \overset{H_0}{\sim} \chi_k^2$, где k размер вектора $\boldsymbol{m}_u$
$LM \overset{H_1}{\sim} \chi_k^2(\gamma)$ [1], где γ — параметр нецентральности, задаваемый в следующем виде:

$$\gamma = \boldsymbol{\xi}_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\xi}_m = m \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} = m\gamma^0,$$

где $\boldsymbol{\xi}_m = \mathbb{E}\boldsymbol{s}_m$, $\boldsymbol{\Sigma}_m = \mathbb{E}[\boldsymbol{s}'_m \boldsymbol{s}_m^T]$.
Также,

$$\gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma)$$

Тогда

$$m^* = \frac{\gamma^*}{\gamma_0}$$

Рассмотрим статистику логарифма отношения правдоподобий:

$$LR = 2\left(l\left(\mathfrak{D}, \hat{\mathbf{w}}\right) - l\left(\mathfrak{D}, \hat{\mathbf{w}}^0\right)\right)$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ — оценка максимума правдоподобия, $\hat{\mathbf{w}}^0 = [\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0]$ — вектор параметров модели, который максимизирует правдоподобие при фиксированном $\mathbf{m}_u^0$.

$LR \overset{H_0}{\sim} \chi_k^2$, где k размер вектора $\boldsymbol{m}_u$

$LR \overset{H_1}{\sim} \chi_k^2$ [2], где γ — параметр нецентральности:

$$\gamma = m\Delta^*, \quad \Delta^* = \mathrm{E}\left[2a^{-1}(\phi)\left\{(\theta - \theta^*)\nabla b(\theta) - b(\theta) + b(\theta^*)\right\}\right],$$

где параметры θ и θ^* считаются по параметрам $\mathbf{w} = [\mathbf{w}_u, \mathbf{w}_v]$ и $\mathbf{w}^* = [\mathbf{w}_u^0, \mathbf{w}_v^*]$ соответственно, вектор параметров $\mathbf{w}_v^*$ задается как решение следующего уравнения:

$$\lim_{m\to\infty} m^{-1}\mathrm{E}\left[\frac{\partial l\left(\mathfrak{D}, \left[\mathbf{m}_u^0, \mathbf{w}_v\right]\right)}{\partial \mathbf{w}_v}\right] = 0.$$

Тогда при заданных α и β получаем оценку размера выборки m^*:

$$m^* = \frac{\gamma^*}{\Delta^*}, \quad \gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma)$$

Вместо обычой логистической и линейной регрессии рассмотрим байесовскую.
Апостериорное распределение:

$$p(w|\mathfrak{D}) = \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, \boldsymbol{V})$$

где $\boldsymbol{V} = I^{-1}(\mathfrak{D}, \hat{\boldsymbol{w}}) = -(\nabla\nabla L(\mathfrak{D}, \hat{\boldsymbol{w}}))^{-1}$ — матрица, обратная матрице Фишера
Для линейной регрессии $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
Для логистической регрессии воспользуемся аппроксимацией Лапласа в точке МП $\hat{\boldsymbol{w}} = argmax_{\boldsymbol{w}} L(\mathfrak{D}, \boldsymbol{w})$

Основная идея

Вычисляется апостериорное распределение для разных подвыборок с возрастающим размером до тех пор, пока не будет выполнено некоторое условие
Условие оптимальности: Апостериорное распределение достаточно локализовано. Локализация может быть описана тремя способами (среднее покрытие, средняя длина, средняя дисперсия).

Критерий средней апостериорной дисперсии

Пусть $\mathrm{D}\hat{\mathbf{w}}$ — Дисперсия апостериорного распределения вектора параметров $\hat{\mathbf{w}}$. Тогда $\mathrm{E}_{\mathfrak{D}_m}\mathrm{D}\left[\hat{\mathbf{w}}|\mathfrak{D}_m\right]$ — матожидание этой дисперсии по всем выборкам размера m из одного распределения
Размер выборки m^* находится из условия:

$$\forall m \geq m^* \ \mathrm{E}_{\mathfrak{D}_m}\mathrm{D}\left[\hat{\mathbf{w}}|\mathfrak{D}_m\right] \leq l$$

Критерий среднего покрытия

Пусть $A(\mathfrak{D}) \subset \mathbb{R}^n$ некоторое множество значений параметров модели $\mathbf{w}$: $A(\mathfrak{D}) = \{\mathbf{w} : ||\mathbf{w} - \hat{\mathbf{w}}|| \leq l\}$, где l — заданный радиус шара.
Критерий для определения m^* :

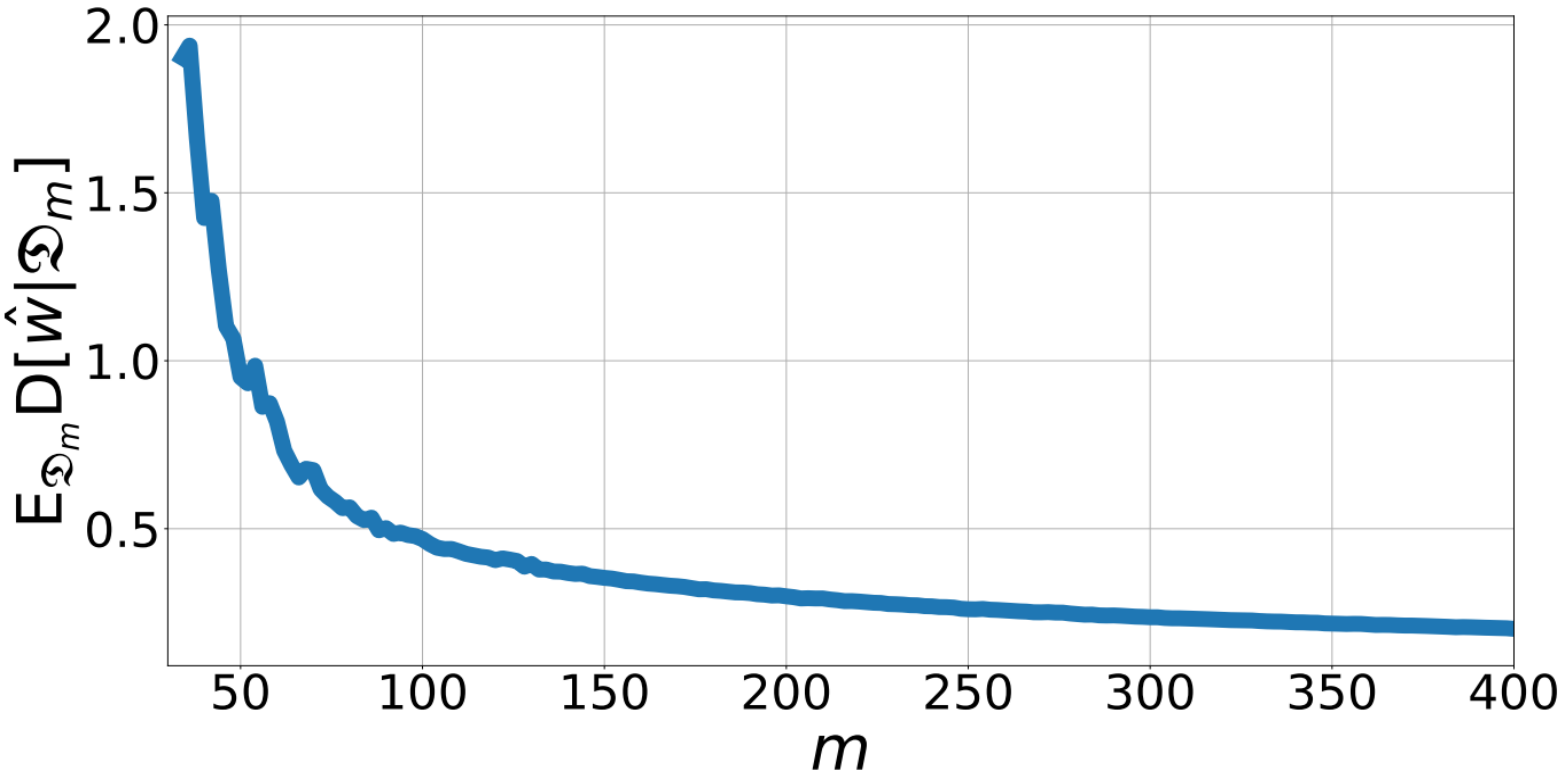
$$\forall m \geq m^* \ \mathrm{E}_{\mathfrak{D}_m}\mathrm{P}\left\{\mathbf{w} \in A(\mathfrak{D}_m)\right\} \geq 1 - \alpha$$

Критерий средней длины

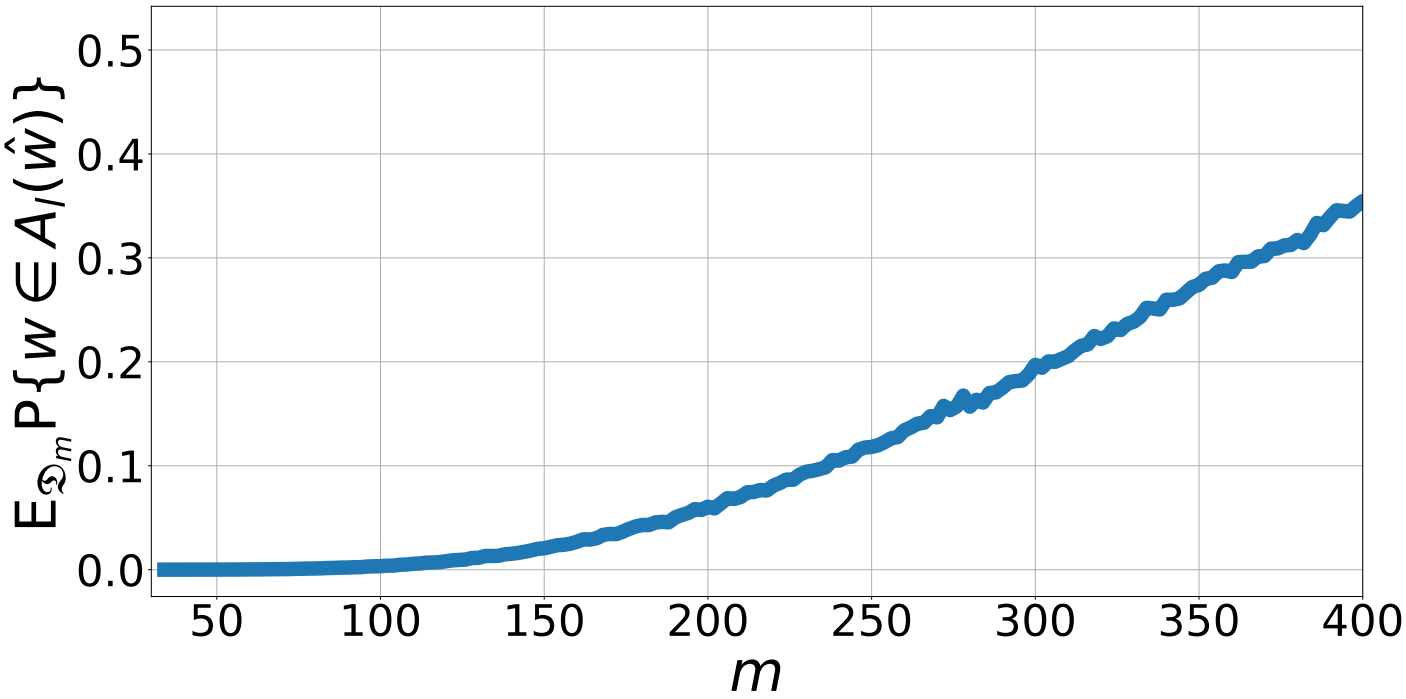
Аналогично предыдущему, но $A(\mathfrak{D})$ определяется как $P(A(\mathfrak{D})) = 1 - \alpha$
Пусть r_m радиус шара $A(\mathfrak{D}_m)$. Критерий для определения m^* :

$$\forall m \geq m^* \ \mathrm{E}_{\mathfrak{D}_m} r_m \leq l$$

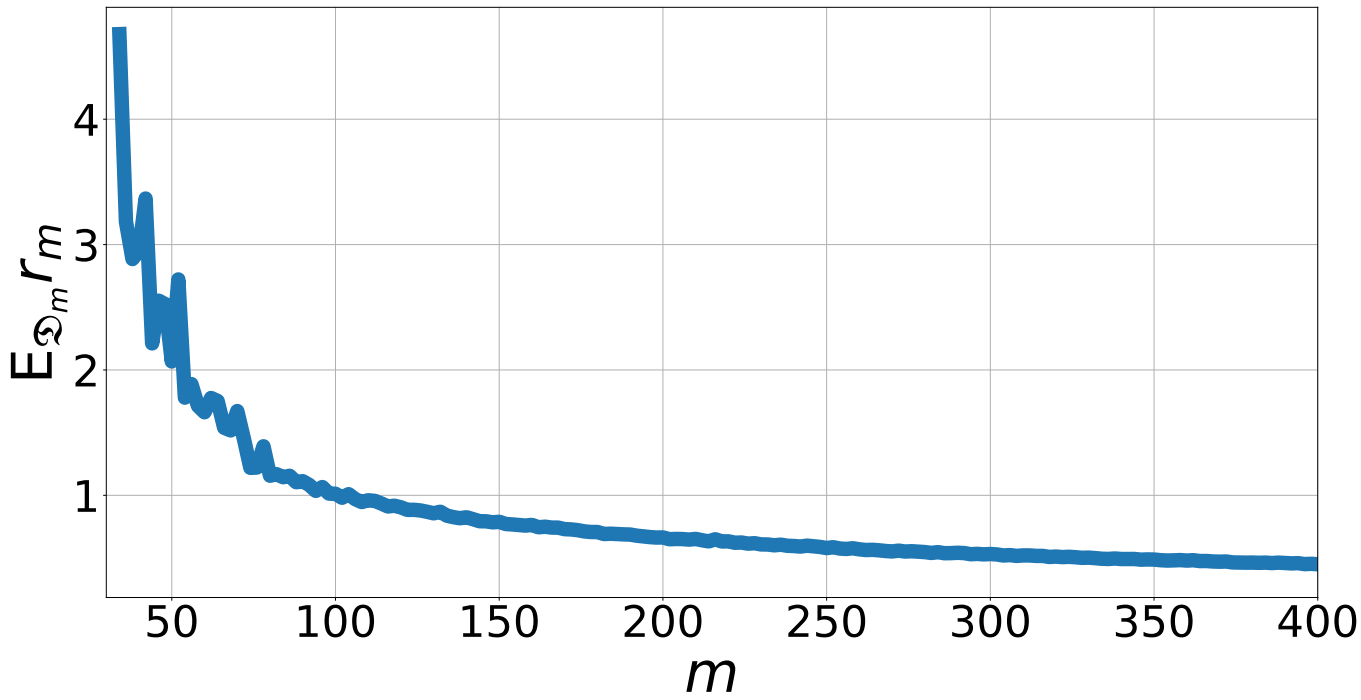
Локализация апостериорного распределения параметров $p(\hat{\boldsymbol{w}}|\mathfrak{D}_m)$ как уменьшение средней дисперсии $\mathrm{E}_{\mathfrak{D}_m}\mathrm{D}\left[\hat{\mathbf{w}}|\mathfrak{D}_m\right]$ по бутстрепным выборкам.



Локализация апостериорного распределения параметров $p(\hat{\boldsymbol{w}}|\mathfrak{D}_m)$ как увеличение средней вероятности попадания вектора параметров в шар радиуса l. Усреднение производится по бутстрепным выборкам.

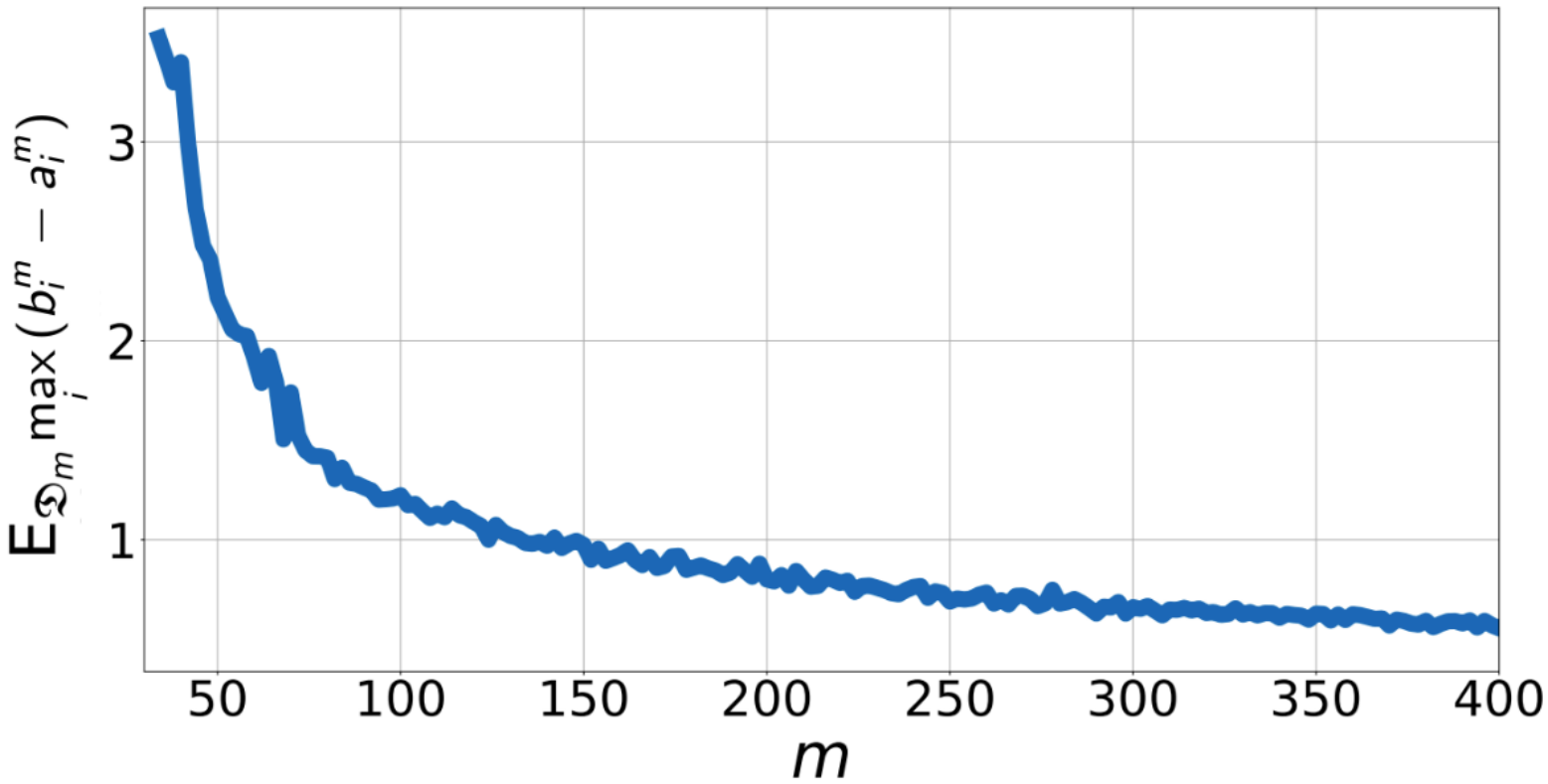


Локализация апостериорного распределения параметров $p(\hat{\boldsymbol{w}}|\mathfrak{D}_m)$ как среднего уменьшение радиуса 0.95-доверительного шара по бутстрепным выборкам.



По выборке размера m построим квантильные бутстрепные доверительные интервалы $(a_1^m, b_1^m), (a_2^m, b_2^m), ..., (a_n^m, b_n^m)$ уровня значимости α для каждого параметра модели.
Оптимальный объём выборки m^* находится из условия малой длины этих интервалов:

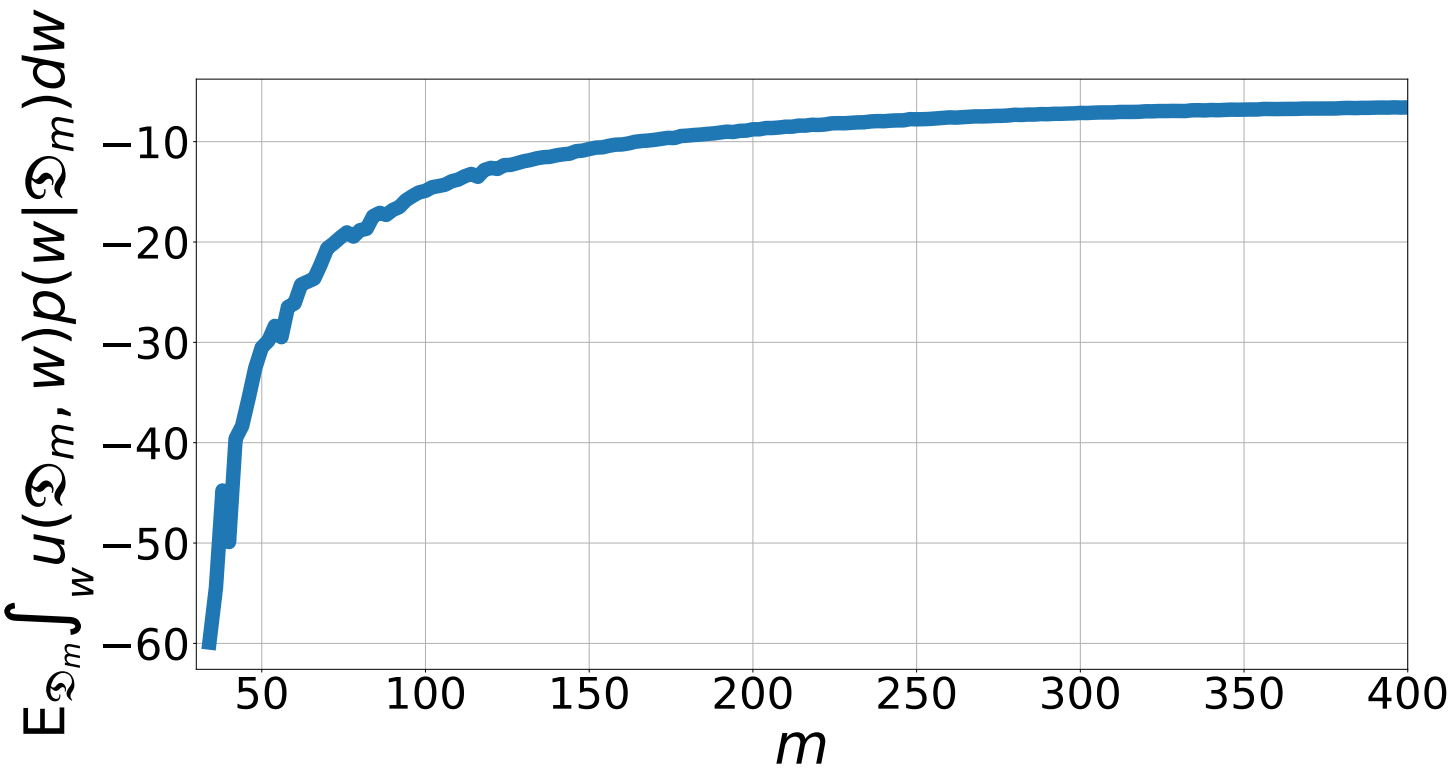
$$m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l.$$



Введём функцию полезности модели $u(\mathfrak{D}_m)$ (в данном случае $u(\mathfrak{D}_m) = lnL(\mathfrak{D}_m, \hat{\boldsymbol{w}})$) и будем искать

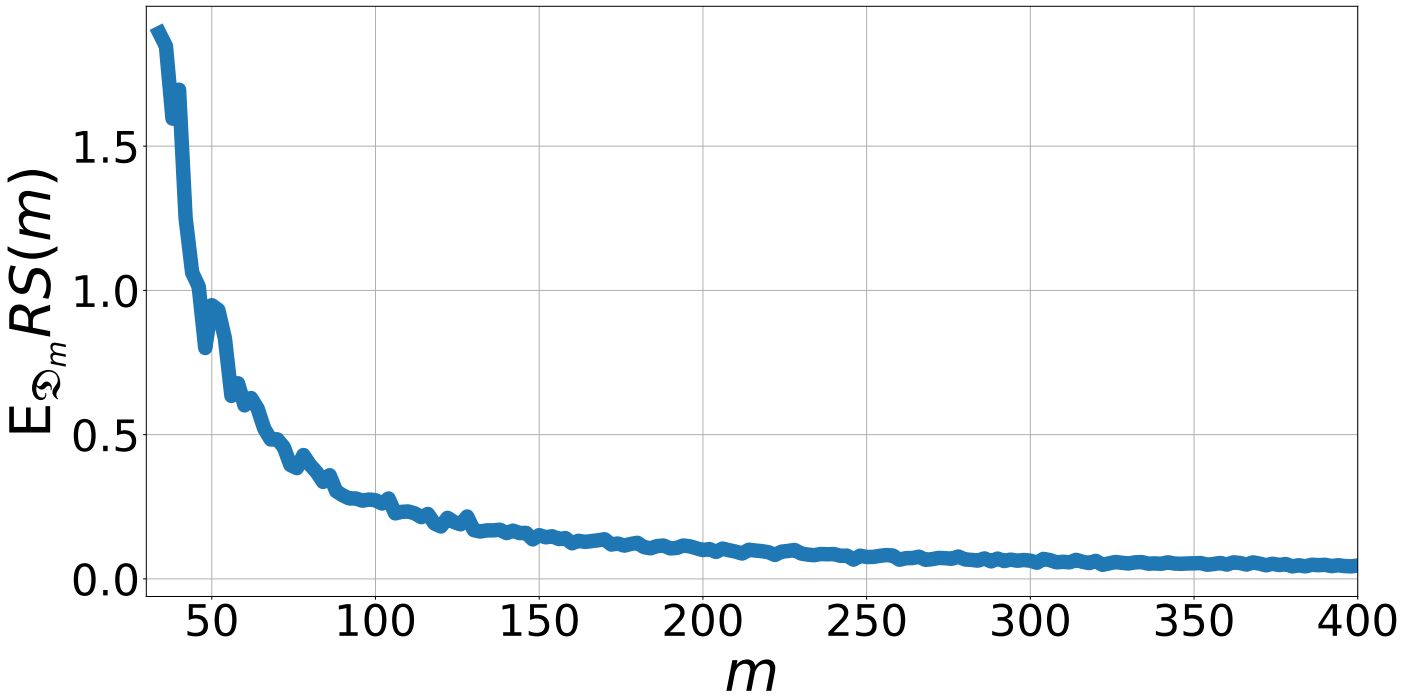
$$m^* = \arg\max_m(\mathbb{E}_{\mathfrak{D}_m}[u(\mathfrak{D}_m)] - cm),$$

где c - коэффициент штрафа за размер выборки

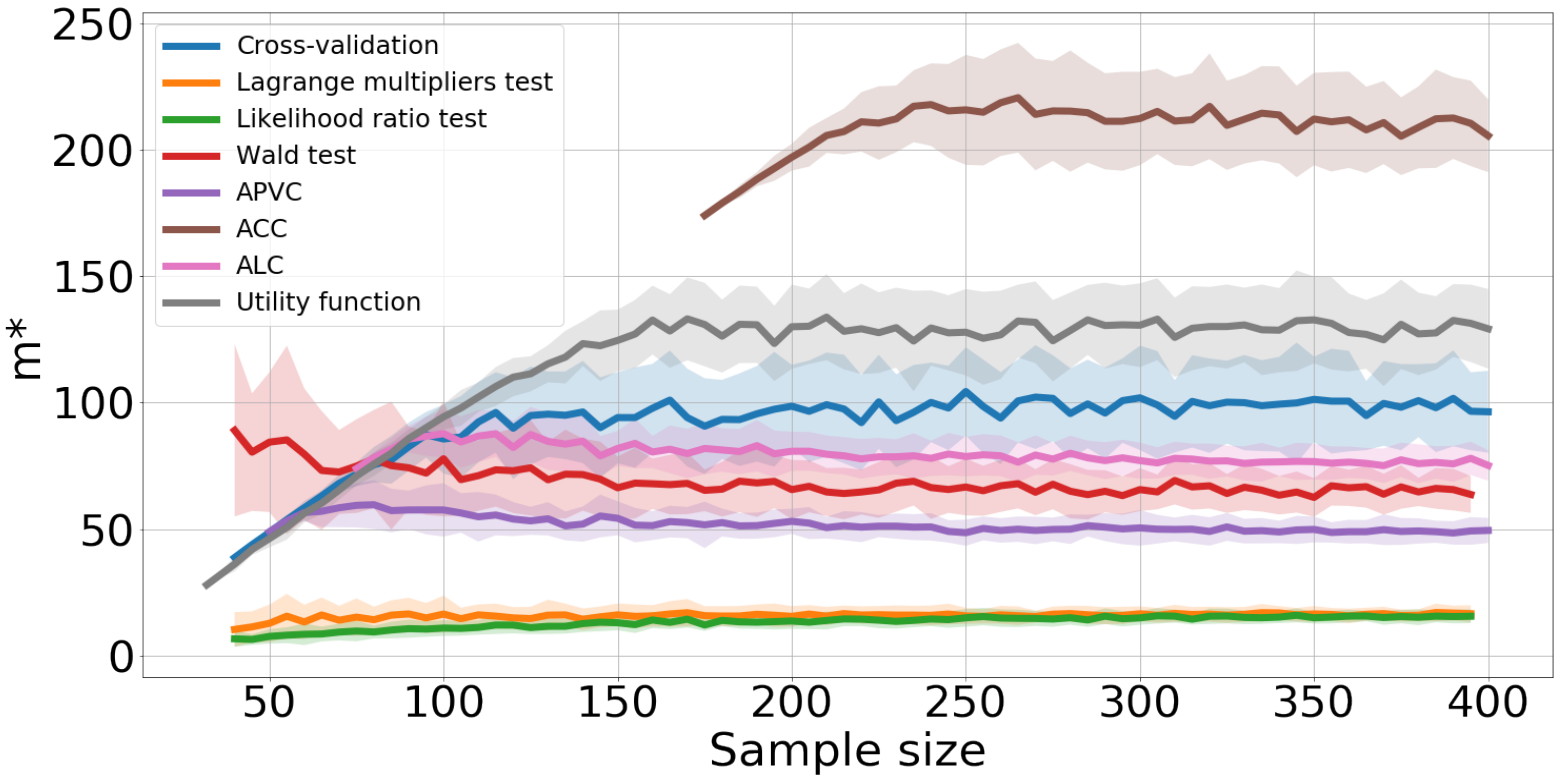


Размер выборки m^* определяется так что для любого $m \geq m^*$ выполнено $RS(m) \geq 1 - \varepsilon$, где ε выбирается малым.

При этом $RS(m) = \ln \frac{L(\mathfrak{D}_{\mathcal{L}(m)}, \hat{\boldsymbol{w}})}{L(\mathfrak{D}_{\mathcal{T}(m)}, \hat{\boldsymbol{w}})}$, где $\mathcal{T}, \mathcal{L}$ — обучающая и тестовая выборки для подвыборки размера m.



Дисперсия при каждом m вычисляется по разным подвыборкам размера m. Все графики выходят на константу.



Выводы

- Исследовано поведение различных методов оценки оптимального объёма выборки
- Описана методология оценки оптимального объёма выборки
- Реализовано программное обеспечение для ретроспективной и предсказательной оценки оптимального объёма выборки


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


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


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.


E. Demidenko Sample size determination for logistic regression revisited // Statist. Med., 2007. Vol. 26. P. 3385–3397.


A. Motrenko and V. Strijov and W. Weber Sample size determination for logistic regression // Journal of Computational and Applied Mathematics, 2014. Vol. 255. P. 743–752.


Maher Qumsiyeh Using the bootstrap for estimation the sample size in statistical experiments // Journal of modern applied statistical methods, 2002. Vol. 8(3). P. 305–321.


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.


D. V. Lindley The choice of sample size // The Statistician, 1997. Vol. 46. P. 129–138.