Оценка необходимого объема выборки пациентов при прогнозировании сердечно-сосудистых заболеваний*

А. П. Мотренко
pastt.petrovna@gmail.com
Московский физико-технический институт, ФУПМ, каф. "Интеллектуальные системы"

В работе описан алгоритм классификации пациентов, перенесших инфаркт и имеющих предрасположенность к инфаркту. Признаками для определения состояния пациента служат измерения концентрации белков в крови. Решается задача оценки параметров функции регрессии и выбора признаков в логистической регрессии. Предполагается, что объем данных недостаточен, поэтому в работе предлагается способ оценки необходимого объема выборки.

Ключевые слова: логистическая регрессия, выбор признаков, оценка объема выборки, прогноз предрасположенности к инфаркту.

Bayesian sample size estimation for logistic regression*

A. P. Motrenko
Moscow Institute of Physics and Technology

The problem of sample size estimation is important in the medical applications, especially in the cases of expensive measurements of immune biomarkers. The paper describes the problem of logistic regression analysis including model feature selection and includes the sample size determination algorithms, namely methods of univariate statistics, logistic regression, cross-validation and Bayesian inference. The authors, treating the regression model parameters as the multivariate variable, propose to estimate sample size using the distance between parameter distribution functions on cross-validated data sets.

Keywords: *logistic regression, sample size, feature selection, Bayesian inference, Kullback-Leibler divergence.*

Введение

Решается задача логистической регрессии [1], в основе которой лежит предположение о биномиальном распределении независимой переменной, и оцениваются параметры функции регрессии [2, 3].

Предполагается, что число измеряемых признаков избыточно. Требуется отыскать оптимальный набор признаков, эффективно разделяющий классы. Признаки в логистической регрессии как правило выбираются с помощью шаговой регрессии [4, 5]. В данной работе используется полный перебор, так как он дает экспертам гарантию, что рассмотрены все возможные сочетания признаков при выборе модели. При этом экспертами вводились ограничения на сложность модели. Задача выбора признаков поставлена с использованием площади под ROC-кривой [6] в качестве внешней функции ошибки.

Задача классификации сопряжена с оценкой минимального объема выборки, достаточного для проведения классификации. Для этого используются следующие методы:

1. Метод доверительных интервалов, в основе которого лежат статистические методы [7].

Научный руководитель В. В. Стрижов

2. Метод оценки объема выборки в логистической регрессии, предложенный Демиденко [8, 9]. Этот способ также основан на на методах математической статистики, но в отличие от метода доверительных интервалов, учитывает распределение зависимой переменной и постановку задачи.
3. Метод скользящего контроля, позволяющий оценить необходимый объем выборки с точки зрения контроля над переобучением [10].
4. Сравнение плотностей распределения параметров модели на различных подвыборках с помощью расстояния Кулльбака-Лейблера [12].

При проведении вычислительного эксперимента были использованы данные [13], подготовленные специалистами парижской лаборатории анализа крови «Иммуноклин».

Задача классификации и оценка параметров

Дана выборка $D = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m$, состоящая из m объектов (пациентов), каждый из которых описывается n признаками (биомаркерами) $\mathbf{x}_i \in \mathbb{R}^n$ и принадлежит одному из двух классов $y_i \in \{0, 1\}$. Рассмотрим задачу логистической регрессии. Предполагается, что вектор ответов $\mathbf{y} = [y_1, \ldots, y_m]^T$ — бернуллиевский случайный вектор с независимыми компонентами $y_i \sim \mathcal{B}(\theta_i)$ и плотностью

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i}. \tag{1}$$

Определим функцию ошибки следующим образом:

$$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln \theta_i + (1-y_i)\ln(1-\theta_i). \tag{2}$$

Другими словами, функция ошибки есть логарифм плотности, или функции правдоподобия, со знаком минус. Требуется оценить вектор параметров $\hat{\mathbf{w}}$, доставляющий минумум функции ошибки:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} E(\mathbf{w}). \tag{3}$$

Вероятность принадлежности объекта к одному из двух классов определим как

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

Для оценки параметров, воспользовавшись тождеством

$$\frac{df(\xi)}{d\xi} = f(1-f),$$

вычислим градиент функции $E(\mathbf{w})$:

$$\nabla E(\mathbf{w}) = -\sum_{i=1}^{m}\big(y_i(1-\theta_i) - (1-y_i)\theta_i\big)\mathbf{x}_i = \sum_{i=1}^{m}(\theta_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\theta} - \mathbf{y}),$$

где вектор $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^T$ и матрица $\mathbf{X} = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T\right]^T$ состоит из векторов-описаний объектов.

Параметров оцениваются методом Ньютона-Рафсона. Введем обозначение $\mathbf{\Sigma}$ — диагональная матрица с элементами $\Sigma_{ii} = \theta_i(1-\theta_i)$, $i = 1, \ldots, m$. В качестве начального приближения $\mathbf{w} = [w_1, \ldots, w_n]^T$ вектора $\hat{\mathbf{w}}$ возьмем

$$w_j = \sum_{i=1}^{m} y_i(1-y_i), \quad j = 1, \ldots, n.$$

Оценка параметров $\mathbf{w}_{k+1}$ логистической регрессии (4) на $k+1$-м шаге итеративного приближения имеет вид

$$\mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\theta} - \mathbf{y}) = (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}(\mathbf{X}\mathbf{w}_k - \mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{y})). \tag{5}$$

Процедура оценки параметров повторяется, пока евклидова норма разности $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|_2$ не станет достаточно мала.

Алгоритм классификации имеет вид:

$$a(\mathbf{x}) = \operatorname{sign}\big(f(\mathbf{x}, \mathbf{w}) - c_0\big), \tag{6}$$

где c_0 — задаваемое в (7) пороговое значение (англ. cut-off) функции регрессии (4).

Вычисления качества прогноза. В данной работе для оценки качества прогноза и для выбора признаков используется площадь AUC под кривой ROC, то есть кривой в осях (FPR(ξ), TPR(ξ)), где

$$\mathrm{TPR} = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) = 1][y_i = 1]$$

есть доля объектов выборки, правильно классифицированных в пользу данного класса, и

$$\mathrm{FPR} = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) = 1][y_i = 0]$$

есть доля ошибочно классифицированных в пользу данного класса объектов выборки. Здесь используется обозначение индикаторной функции:

$$[y = 1] = \begin{cases} 1, & y = 1; \\ 0, & y \neq 1. \end{cases}$$

Таким образом, алгоритм тем лучше разделяет классы чем больше значение AUC.

Отыскание параметра c_0 алгоритма классификации. Каждая точка кривой ROC соответствует некоторому значению c_0. В алгоритме (6) используется то значение c_0, которое соответствует наибольшему расстоянию от отрезка [(0,0);(1,1)], означающего отказ от принятия решения о классификации, до кривой ROC:

$$\hat{\sigma}_0 = \arg\max_{\xi\in[0,1]} \big\|\big(\mathrm{TPR}(\xi), \mathrm{FPR}(\xi)\big) - (\xi, \xi)\big\|_1. \tag{7}$$

Последнее выражение включает вычисление значения функционала качества, и как следствие, вычисление выражения (6) и итеративную оценку параметров (5).

Выбор признаков в задаче классификации

Введем обозначения $\mathcal{A}$ — некоторое подмножество индексов признаков, $\mathcal{A} \subseteq \mathcal{J} = \{1, \ldots, n\}$ и $\hat{\mathcal{A}}$ — оптимальный набор индексов. Обозначим $\mathbf{X}_\mathcal{A}$ множество столбцов-признаков матрицы $\mathbf{X}$, заданное набором $\mathcal{A}$, и $\mathbf{w}_\mathcal{A}$ — соответствующие им параметры. Рассмотрим задачу выбора признаков как задачу максимизации:

$$\hat{\mathcal{A}} = \arg\max_{\mathcal{A} \subseteq \mathcal{I}} S(\mathcal{A}) \text{ при условии } |\mathcal{A}| = \text{const}. \tag{8}$$

В задаче использована площадь под кривой $S(\mathcal{A}) \equiv S(\mathbf{X}_\mathcal{A}, \hat{\mathbf{w}}_\mathcal{A}, \hat{\sigma}_0, \mathbf{y})$, значение которой вычислено для набора индексов признаков $\mathcal{A}$, а параметры $\hat{\mathbf{w}}_\mathcal{A}$ и c_0 получены в результате решения задач (3) и (7).

Набор признаков отыскивается путем полного перебора. Такой подход возможен благодаря сравнительно небольшому количеству признаков в данной задаче и диктуется требованиями экспертов.

Так как количество признаков в искомом наборе $\mathcal{A}$ неизвестно, множество индексов объектов $\mathcal{I}$ разбивается случайным образом на два подмножества, $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$, обучающее и тестовое. Параметры $\mathbf{w}$ оцениваются на подвыборке $D_\mathcal{L}$, а качество прогноза вычисляется на подвыборке $D_\mathcal{T}$. Максимальное число признаков при решении задачи фиксировано экспертами: $|\mathcal{A}|$ не должна превышать четырех. Наборы признаков, полученные в результате решения задачи (8), будем называть оптимальным, а сами признаки — наиболее информативными.

Оценка объема выборки

Данные, использованные при проведении вычислительного эксперимента, содержат признаковые описания пациентов, принадлежащих одному из классов: больные, перенесшие инфаркт или имеющие предрасположенность к инфаркту. В качестве признаков (биомаркеров) используются концентрации белков и их соединений, абсорбированные на поверхности кровяных телец. В классах содержится четырнадцать и сорок объектов соответственно. При таком объеме данных возникает задача оценки минимального объема выборки m^*, необходимого для получения статистически достоверных результатов классификации. В данном разделе рассмотрены четыре способа оценки объема выборки. Результаты оценки объема выборки описаны и проанализированы в разделе «Вычислительный эксперимент».

Метод доверительных интервалов. Рассмотрим выборку, в который каждый объект описывается одним признаком $D = \{(x_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$. Пусть $\Delta = \bar{x} - \mu$ — разница между средним арифметическим

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

и известным математическим ожиданием μ. При известном среднеквадратическом отклонении σ случайная величина принадлежит страндартному нормальному распределению

$$Z = \frac{\bar{x} - \mu}{\sigma} \sqrt{m} = \frac{\Delta}{\sigma} \sqrt{m} \sim \mathcal{N}(0, 1). \tag{9}$$

Тогда

$$\Delta = z_{\alpha/2} \frac{\sigma}{\sqrt{m}},$$

где $z_{\alpha/2}$ таково, что $P\left\{|Z| \geqslant z_{\alpha/2}\right\} = \alpha$. Отсюда получаем формулу для оценки размера выборки

$$m^* = \left(\frac{z_{\alpha/2}\sigma}{\Delta}\right)^2. \tag{10}$$

При $m \geqslant 30$ можно пользоваться предположением о нормальности случайной величины Z, если величины x_i распределены ненормально, а также при неизвестном σ^2, заменив его в выражении (9) на

$$s = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2}.$$

Однако в случае $m \leqslant 30$ для использования этой формулы необходимо, чтобы случайные величины x_i были распределены нормально; кроме того, среднеквадратическое отклонение σ должно быть известно.

В данной работе рассматривается многопризнаковая задача, однако в предположении, что все признаки из наиболее информативных наборов взаимно независимы, формула (10) верна для каждого из них. Вычисляя объем выборки для различных признаков, будем получать различные значения. Для получения общей оценки можно взять среднее или наибольшее из них. При таком подходе можно получить лишь грубую оценку, так как более правдоподобна альтернативная гипотеза о том, что распределение признаков в выборке представляет собой смесь из двух нормальных распределений:

$$x_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{с вероятностью } p_i; \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{с вероятностью } 1 - p_i, \end{cases} \tag{11}$$

где параметр p_i задан с помощью (4)

Метод оценки объема выборки в логистической регрессии. Фиксируем некоторое множество $\mathcal{A}$ индексов признаков, используемых для получения прогноза. Для каждого из признаков, вошедших в этот набор, можно оценить объем выборки, необходимый чтобы включить этот признак в набор. Для этого рассмотрим нулевую гипотезу вида

$$H_0 : w_j = 0,\ j \notin \mathcal{A},$$

где w_j — j-тая компонента вектора параметров $\mathbf{w}$ логистической регрессии. Таким образом, нулевая гипотеза заключается в предположении, что j-тый признак не включается в модель. Оценив вектор параметров при нулевой гипотезе, получим $\mathbf{w}_{\mathcal{A}}$, а при принятии альтернативной гипотезы — $\mathbf{w}_{\mathcal{A}^*}$, где множество $\mathcal{A}^*$ получается из $\mathcal{A}$ добавлением к нему индекса j. Тогда нулевая и альтернативная гипотезы могут быть переформулированы относительно параметров θ_i бернуллиевского распределения $\mathcal{B}(\theta)$ и представлены в виде

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathcal{A}}, H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathcal{A}^*}.$$

При этом неважно, какие именно значения принимают θ_i в каждом случае, интерес представляет только пороговое значение функции регрессии θ_0. Окончательно сформулируем гипотезы в виде:

$$H_0 : 1 - {}_0 = p_0, H_1 : 1 - {}_0 = p_1.$$

Выберем в качестве тестовой статистики

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / m}}, \quad \hat{p} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

где $\hat{p}$ — оценка максимального правдоподобия для параметра θ. При принятии альтернативной гипотезы статистика Z имеет нормальное распределение

$$Z \sim \mathcal{N}\left(p_1 - p_0, \sqrt{\frac{p_1 q_1}{p_0 q_0}}\right).$$

Тогда величина

$$Z\sqrt{\frac{p_0 q_0}{p1 q_1}} + \frac{p_0 - p_1}{\sqrt{p_1 q_1 / m}} = \sqrt{\frac{p_0 q_0}{p_1 q_1}}\left(Z + \frac{p_0 - p_1}{\sqrt{p_0 q_0}}\sqrt{m}\right) \sim \mathcal{N}(0,1).$$

Выбрав уровень значимости α и мощность критерия $1-\beta$, запишем выражение для мощности

$$1 - \beta = P\{|Z| > Z_{\alpha/2}|H_1\} = \Phi\left(\sqrt{\frac{p_0 q_0}{p_1 q_1}}\left(Z_{\alpha/2} + \frac{p_0 - p_1}{\sqrt{p_0 q_0 / m}}\right)\right).$$

Тогда необходимый объем выборки вычисляется по формуле

$$m^* = \frac{p_0 q_0 \left(Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\frac{p_1 q_1}{p_0 q_0}}\right)^2}{(p_1 - p_0)^2}. \tag{12}$$

Заметим, что вычисленный объем выборки зависит от номера признака, относительно которого сформулирована нулевая гипотеза.

Скользящий контроль. Метод скользящего контроля позволяет оценить необходимый объем выборки с точки зрения контроля над переобучением. При таком подходе производится разбиение выборки на две непересекающиеся подвыборки: обучающую и тестовую. Пусть $\mathcal{I} = \{1, \ldots, m\}$ — множество индексов объектов выборки, разобьем его на два непересекающихся подмножества $\mathcal{I} = \mathcal{L} \bigsqcup \mathcal{T}$. Тогда обучающей выборкой назовем $D_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{L}$, а тестовой — $D_{\mathcal{T}} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{T}$. Фиксируем набор признаков с индексами $\mathcal{A}$. Уменьшение функционала качества $S(\mathcal{A}, D_{\mathcal{T}})$ на тестовой выборке по сравнению с его значением $S(\mathcal{A}, D_{\mathcal{L}})$ на обучающей выборке свидетельствует о наличии переобучения. Переобучением назовем отношение

$$\mathrm{RS}(m) = \frac{S(\mathcal{A}, D_{\mathcal{T}(m)})}{S(\mathcal{A}, D_{\mathcal{L}(m)})}. \tag{13}$$

В этом случае модель f хорошо описывает данные, на которых она была настроена, но плохо приближает тестовую выборку. Переобучение может являться следствием недостаточного объема выборки. Чтобы оценить необходимый объем выборки, будем последовательно наращивать объем выборки m, производя разбиение на обучение и контроль в заданном отношении: $|\mathcal{T}(m)|/|\mathcal{L}(m)| = \mathrm{const} \leqslant 0,5$. Для каждого значения m будем вычислять отношение (13). При увеличении объема выборки оно стремится к единице. Будем считать, что объем выборки m^* достаточен, если начиная с него величина RS не меньше чем заданное $1 - \varepsilon_1$.

Оценка объема выборки с использованием расстояния Кулльбака-Лейблера. Предлагаемый подход основан на вычислении расстояния между функциями распределения параметров регрессионной модели. Рассмотрим некоторое множество индексов объектов $\mathcal{B}_1 \in \mathcal{J}$, а также множество $\mathcal{B}_2 \in \mathcal{J}$, такое что:

$$|\mathcal{B}_1 \setminus \mathcal{B}_2 \cup \mathcal{B}_2 \setminus \mathcal{B}_1| \leqslant 2.$$

Таким образом, множество $\mathcal{B}_2$ может быть получено из $\mathcal{B}_1$ путем удаления, добавления или замены одного элемента. Оценивая параметры на различных подвыборках, будем получать различные результаты. На рисунке 1 продемонстрировано, как меняется положение разделяющей гиперплоскости, определяемой выражением

$$\mathbf{x}^T\mathbf{w} = \ln(\frac{c_0}{1-c_0})$$

при добавлении в выборку двух элементов. Если объем выборки $D_{\mathcal{B}_1}$ достаточно велик,

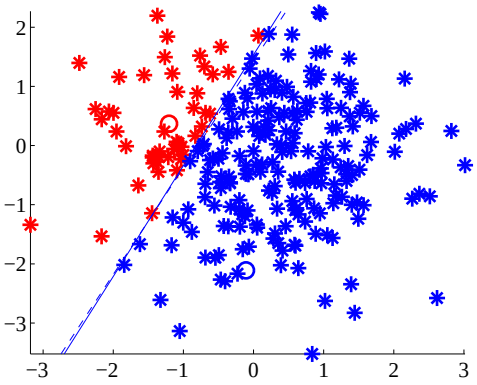


Рис. 1. Два класса, разделенные гиперплоскостью. Пунктирной линией обозначено положение гиперплоскости после того как два новых объекта (выделенных окружностями), были добавлены в выборку.

небольшое изменение ее состава $D_{\mathcal{B}_2}$ не должно приводить к существенному изменению параметров. Простейший способ сравнивать параметры на различных подвыборках — с помощью

$$||\mathbf{w}_1 - \mathbf{w}_2|| = \sqrt{\sum_{i=1}^{|\mathcal{A}|}\left(w_i^1 - w_i^2\right)^2}.$$

Предлагается сравнивать функции распределения параметров модели на подвыборках $D_{\mathcal{B}_1}$ и $D_{\mathcal{B}_2}$ с помощью расстояния Кулльбака-Лейблера.

Рассмотрим модель (4) и предположение о распределении случайной величины y_i (1). Зафиксировав выборку D и модель $f_{\mathcal{A}} = f(X_{\mathcal{A}}^T\mathbf{X})$, перепишем (1) в виде

$$p(\mathbf{y}|X, \mathbf{w}, f_{\mathcal{A}}) \equiv p(D|\mathbf{w}, f_{\mathcal{A}}) = \prod_{i=1}^{m}\theta_i^{y_i}(1-\theta_i)^{1-y_i}. \tag{14}$$

Предположим также, что вектор параметров $\mathbf{w}$ регрессии имеет нормальное распределение $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2 I_{|\mathcal{A}|})$ с плотностью

$$p(\mathbf{w}|f_{\mathcal{A}}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|\mathcal{A}|}{2}}\exp(-\frac{\alpha}{2}||\mathbf{w}-\mathbf{w}_0)||^2), \tag{15}$$

где $\alpha^{-1} = \sigma^2$, а $I_{|\mathcal{A}|}$ — единичная матрица размерности $|\mathcal{A}|$.

Найдем плотность распределения $p(\mathbf{w}|D, \alpha, f_{\mathcal{A}})$ параметров модели, воспользовавшись формулой Байеса

$$p(\mathbf{w}|D, \alpha, f_{\mathcal{A}}) = \frac{p(D|\mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|\alpha, f_{\mathcal{A}})}{p(D|\alpha, f_{\mathcal{A}})}, \tag{16}$$

где $p(D|\mathbf{w}, f_{\mathcal{A}})$ — правдоподобие данных, $p(\mathbf{w}|\alpha, f_{\mathcal{A}})$ — задаваемая априорно плотность распределения параметров модели. В выражении (16) нормировочный множитель $p(D|\alpha, f_{\mathcal{A}})$ определяется выражением

$$p(D|\alpha, f_{\mathcal{A}}) = \int p(D|\mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|\alpha, f_{\mathcal{A}})d\mathbf{w}.$$

Подставив (14) и (15) в (16) и обозначив $Z(\alpha) = p(D|\alpha, f_{\mathcal{A}})$, получим

$$p(\mathbf{w}|D, f_{\mathcal{A}}) = \frac{p(y|\mathbf{x}, \mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|f_{\mathcal{A}}, \alpha)}{Z(\alpha)} =$$
$$= \frac{\alpha^{\frac{|\mathcal{A}|}{2}}}{(2\pi)^{\frac{|\mathcal{A}|}{2}} Z(\alpha)} \exp(-\frac{\alpha}{2}||\mathbf{w} - \mathbf{w}_0)||^2) \prod_{i=1}^{m} \theta_i^{y_i}(1 - \theta_i)^{1-y_i},$$

где $Z(\alpha)$ — нормировочный множитель.

Пусть имеются две выборки, $D_{\mathcal{B}_1}$ и $D_{\mathcal{B}_2}$. Обозначим соответствующие апостериорные распределения $p_1(\mathbf{w}) \equiv p(\mathbf{w}|D_{\mathcal{B}_1}, \alpha, f_{\mathcal{A}})$ и $p_2(\mathbf{w}) \equiv p(\mathbf{w}|D_{\mathcal{B}_2}, \alpha, f_{\mathcal{A}})$. Для оценки сходства плотностей распределения параметров вычислим расстояние Кулльбака-Лейблера между ними

$$D_{\mathrm{KL}}(p_1, p_2) = \int_{\mathbf{w}\in\mathcal{W}} p_1(\mathbf{w}) \frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} d\mathbf{w}. \tag{17}$$

Используя в качестве меры изменения распределения $p(\mathbf{w}|D, f_{\mathcal{A}})$ расстояние Кулльбака-Лейблера, оценим необходимый объем выборки. Для этого будем N раз случайным образом удалять из выборки по одному объекту, уменьшая ее размер, и вычисляя каждый раз плотность распределения вектора $\mathbf{w}$ с помощью (15). Затем посчитаем расстояние (17) между «соседними» распределениями, т.е. между функциями распределения параметров, которые оценивались на подвыборках, отличающихся друг от друга только одним объектом. Проведем эту процедуру N раз и усредним полученные расстояния. Считаем объем выборки m^* достаточным, если начиная с расстояние (17) меняется не больше чем на заранее заданное число ε_2.

Результаты вычислительного эксперимента

Эксперимент на реальных данных. В данном разделе описан вычислительный эксперимент, который проводился на данных лаборатории анализа крови «Иммуноклин». Данные содержат измерения концентрации 20-ти белков и их соединений на поверхности кровяных телец пациентов двух классов, содержащих 31 и 14 объектов соответственно. В таблице 2 приведен список исследуемых биомаркеров с их порядковыми номерами.

Таблица 1. Результаты выбора признаков

$\mathcal{A}$	$S(\mathcal{A})$
K, L , L/P	0.9750
K, L, K/M, K/Q	0.9671
K, L, L/M, L/T/SO	0.9933
K, L, K/M, L/R	0.9867
K, K/M, L/P,	0.9742

В таблице 1 указаны наборы маркеров, доставивших наибольшие значения максимизируемому критерию AUC и сами значения этого критерия. Для исследования были выбраны K лучших наборов.

В данном случае выбрано значение $K = 5$.

Таблица 2. Число вхождений признаков в K оптимальных наборов для каждого признака.

K	L	K/M	L/M	K/N	K/O	L/O	K/P	L/P	K/Q
5	4	3	1	0	0	0	0	2	1

K/R	L/R	L/R/SA	L/T/SA	L/T/SO	U/V	U/W	U/X	U/Y	U/Z
0	1	0	0	1	0	0	0	0	0

Одной из важный практических задач, решаемых в рамках проводимых исследований, является задача снижения стоимости клинического исследования одного пациента, решаемая путем уменьшения числа измеряемых биомаркеров. Предложено измерять только наиболее информативные биомаркеры, выбранные следующим образом. Объединив признаки из всех наборов из колонки «$\mathcal{A}$» таблицы 1, получим множество наиболее информативных признаков $\mathcal{S} = \bigcup_{i=1}^{K} \{\mathcal{A}_i\}$. Для каждого признака подсчитано количество его вхождений в это множество. Таблица 2 показывает число вхождений каждого биомаркера в $\mathcal{S}$.

Оценка необходимого объема выборки

На гистограмме 2 отложены оценки объема выборки m^*, вычисленные по формулам (10) и (12), от признака. Заметим, что нет необходимости проводить усреднение, как это предлагалось в разделе «Метод доверительных интервалов», по всем признакам, так как в модель вошли лишь некоторые из них, остальные являются неинформативными и учитываться не должны.

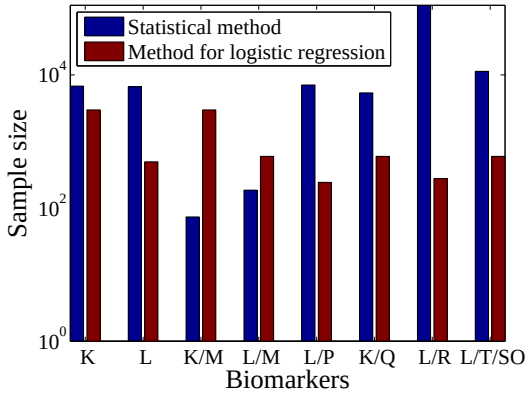


Рис. 2. Оценка объема выборки, полученная с помощью метода доверительных интервалов и с помощью метода для логистической регрессии, для каждого из признаков.

Заметим, что зависимости, изображенные на рисунке 2 носят одинаковый характер: положения максимумов и минимумов практически совпадают. Это происходит от того, что размер выборки, оцененный для j-го признака, зависит от информативности этого признака. В логистической регрессии такие признаки имеют большое по абсолютной величине значение соответствующего элемента вектора параметров w_j. В формуле (12) в знаменателе стоит квадрат разности $(c_0 - \sigma_1)^2$. Чем ближе к нулю величина w_j, тем меньше $(c_0 - \sigma_1)^2$, и больше m^*. Таким образом, наименьшие значения объема выборки соответствуют наиболее информативным признакам, а аномально большие (порядка 10^4 и выше) наблюдаются у признаков, которые в модель не входят — у них w_j близко к нулю.

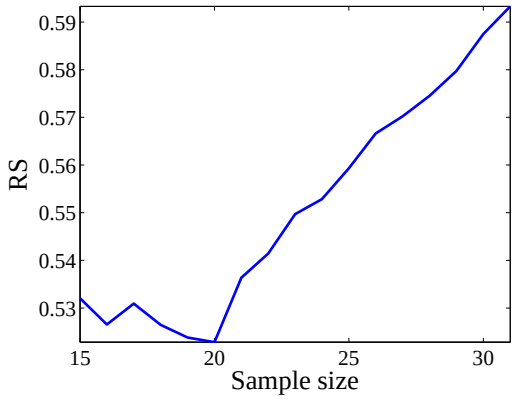
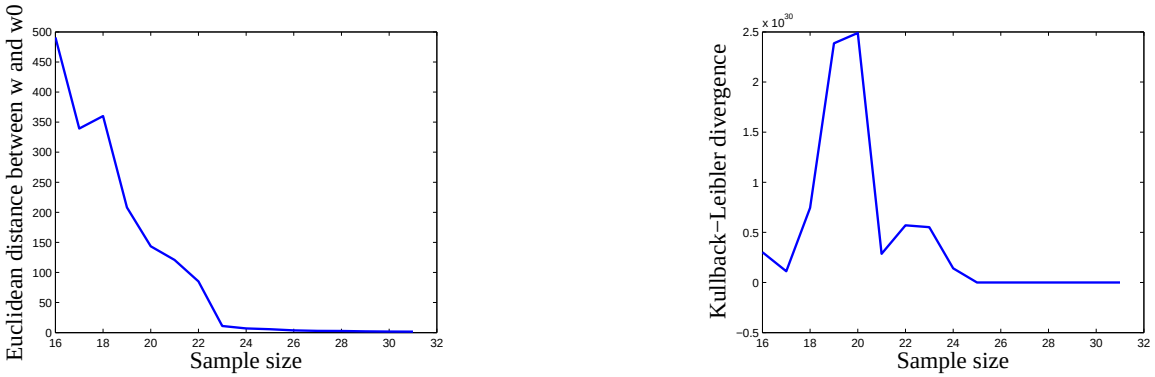


Рис. 3. Оценка объема выборки, полученная с помощью метода Демиденко, для каждого из признаков.



a. b.

Рис. 4. a. Усредненное еклидово расстояние между параметрами модели, $||\mathbf{w}_m - \mathbf{w}_{m+1}||$ b. Усредненное расстояние Кулльбака-Лейблера.

На рисунке 3 изображена зависимость величины $\mathrm{RS}(m)$, определяемой (13) от размера выборки, на которой проводился скользящий контроль. При данном размере выборки функция $\mathrm{RS}(m)$ не успевает выйти на асимптоту, и о дальнейшем поведении функции по рисунку 3 судить нельзя, поэтому метод скользящего контроля дает оценку на необходимый объем выборки $m^* \geqslant 30$.

На рисунке 4 изображены зависимости евклидового расстояния между параметрами и расстояния Кулльбака-Лейблера между их плотностями, усредненного по $N = 100$ разбиениям, от количества объектов в выборке. Видно, что при $m \geqslant 25$ оба графика меняется достаточно плавно. Таким образом, минимальный объем выборки, оцененный с помощью расстояния Кулльбака-Лейблера $m^* \leqslant 30$.

Таблица 3. Результаты оценки необходимого объема выборки с помощью четырех различных методов.

Метод доверительных интервалов	Демиденко	Скользящий контроль	Расстояние Кулльбака-Лейблера
$10^2 - 10^4$	$\sim$50	$\geqslant 30$	$\geqslant 30$

Для сравнения перечисленных методов, приведем таблицу 3, содержащую оценки необходимого объема выборки для каждого из рассмотренных методов. Особенностью использованных при проведениии вычислительного эксперимента данных является небольшой объем выборки, поэтому метод скользящего контроля и расстояние Кулльбака-Лейблера дают лишь нижнюю оценку, так как эти методы больше подходят для оценки достаточного объема выборки, то есть применимы когда объем выборки слишком велик. Метод доверительных интервалов и метод Демиденко дают результаты, численно отличающиеся на порядки, однако качественно схожие друг с другом. Последнее объясняется тем, что объем выборки, оцениваемый этими методами, зависит от информативности признака. Различие можно объяснить грубостью предположений, сделанных в рамках метода доверительных интервалов.

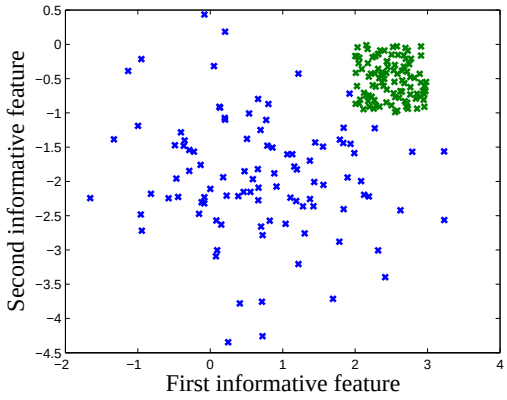


Рис. 5. Распределение данных в пространстве двух информативных признаков.

Эксперимент на синтетических данных. Работа алгоритма также была протестирована на примере с искусственными данными, их структура отображена на рисунке 5. Оба класса имеют по один шумовой и два информативных признака и содержат по 100 объектов.

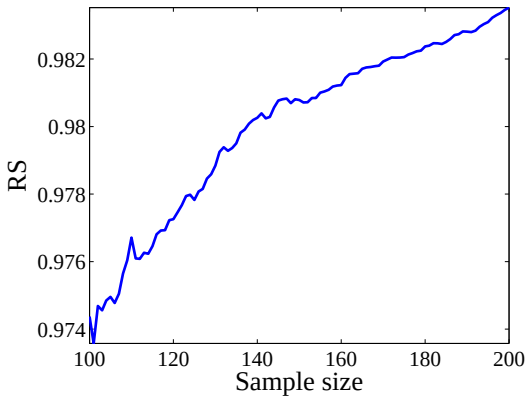


Рис. 6. Зависимость RS(m) при разделении выборки на обучающую и контрольную в отношении 3:1.

Из рисунка 6 видно, что уже начиная с $m = 100$ величина RS(m) меняеется не более чем на 0.001, т.е. можно считать, что $m^* \leqslant 100$.

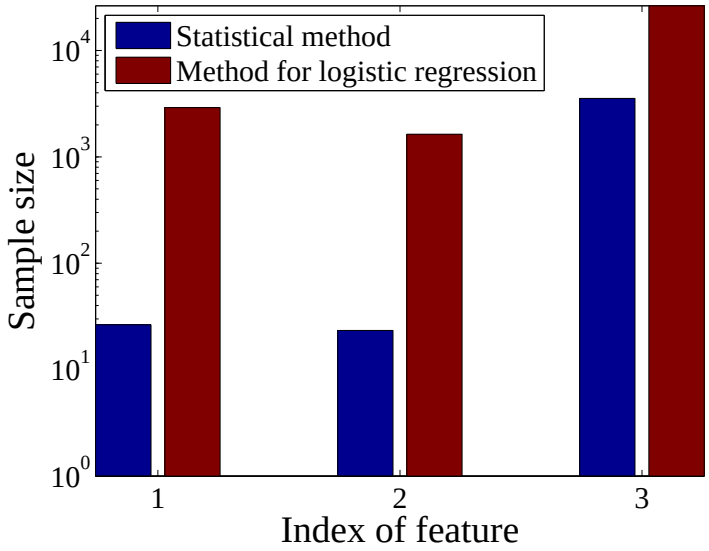


Рис. 7. Оценка объема выборки, полученная с помощью метода доверительных интервалов и с помощью метода для логистической регрессии, для каждого из признаков.

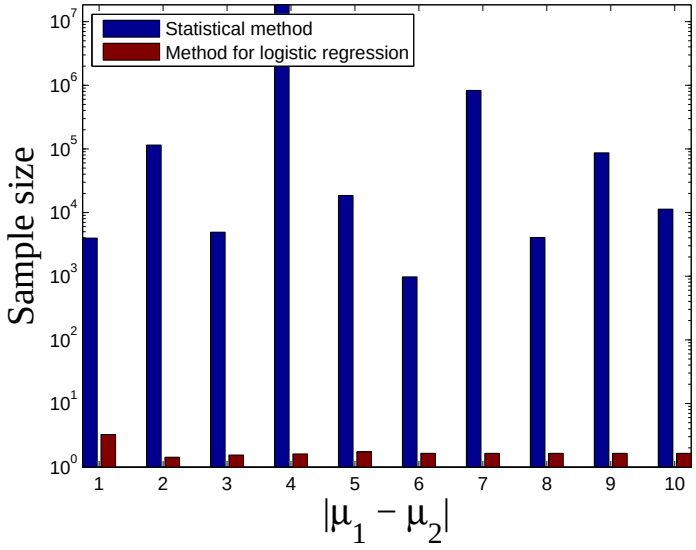


Рис. 8. Оценка объема выборки, полученная с помощью метода доверительных интервалов и с помощью метода для логистической регрессии, для для различных значений $|\mu_1 - \mu_2|$.

На гистограмме 7 отражены результаты оценки m^* методом доверительных интервалов и методом для логистической регрессии. При этом менее точный метод доверительных интервалов дает результат, более близкий к m^*, полученному с помощью методов скользящего контроля и расстояния Кулльбака-Лейблера. Дело в том, что распределение рассматриваемых данных сильно отличается от предполагаемого расределения (11) реальных данных, описанных в разделе 12. Рассмотрим выборку, с признаковым описанием, состоящим из одной случайной величины, распределенной как (11). Меняя расстояние $|\mu_1 - \mu_2|$ между математическими ожиданиями компонент смеси, будем наблюдать за результата-

ми оценки m^*, полученными с помощью методов скользящего контроля и логистической регрессии. Результаты приведены на гистограмме 8

В этом случае метод скользящего контроля дает завышенные результаты, в то время как метод для логистической регрессии оказывается более точен.

Заключение

В работе описан алгоритм прогнозирования вероятности наступления инфаркта пациентов; описан способ оценки параметров и выбора наиболее информативных признаков методом полного перебора. Этот подход возможен благодаря небольшому количеству признаков и дает экспертам гарантию, что выбран оптимальный набор. Описаны способы получения оценки необходимого объема выборки пациентов. Предложен новый метод оценки объема выборки, основанный на вычислении расстояния Кулльбака-Лейблера между плотностями распределения параметров модели, оцениваемыми при различных разбиениях выборки. Все методы протестированы на реальных и синтетических данных.

Литература

[1] *Hosmer D., Lemeshow S.* Applied logistic regression. N. Y.: Wiley, 2000. 375 p.

[2] *Bishop C. M.* Pattern recognition and machine learning. Springer, 2006. 738 p.

[3] *MacKay D. J. C.* Information theory, inference, and learning algorithms. Cambridge University Press, 2003. 628 p.

[4] *Friedman J., Hastie, Tibshirani R.* Additve logistic regression: a statistical way of boosting // The Annals of Statistics. 2000. V. 28, № 2. P. 337–407.

[5] *Madigan D., Rideway G.* Discussion of least square regression. В сб. Efron B. [et al.]. Least Angle Regression // The Annals of Statistics. 2004. V. 32, № 2. P. 465–469.

[6] *Fawcet T.* ROC graphs: notes and practical considerations for researchers // HP Laboratories, 2004. 38 p.

[7] *Реброва О. Ю.* Статистический анализ медицинских данных. Применение прикладного пакета Statistica. М.: Медиасфера, 2002. 312 с.

[8] *Demidenko E.* Sample size determination for logistic regression revisited // Statist. Med. 2007; 26:3385–3397.

[9] *Rosner B.* Fundamentals of biostatistics. Duxbury Press, 1999. 816 p.

[10] *Bos S.* How to partition examples between cross-validation set and fraining set? / Saitama, Japan: Laboratory for information representation RIKEN. 1995. 4 p.

[11] *Amari S., Murata N., Muller K.-R., Finke M., Yang H.H.* Asymptotic statistical theory of overtraining and cross-validation. // IEEE Transactions on Neural Networks, 1997. V. 8, No. 5. P. 985–996.

[12] *Perez-Cruz F.* Kullback-Leibler divergence estimation of continuous distributions // IEEE International Symposium on Information Theory, 2008.

[13] Standart flow cytometry analysis of nondental patients. Paris: ImmunoClin laboratory. 2007. 1 p.