Оценка объема выборки в задачах логистической регрессии

А. П. Мотренко

Научный руководитель: В. В. Стрижов
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

13 июня 2012

Работа посвящена прогнозу вероятности принадлежности пациента к одному из нескольких неупорядоченных классов:

- A_1 Больные, перенесшие инфаркт.
- A_2 Больные, находящиеся в состоянии инфаркта.
- A_3 Пациенты, имеющие предрасположенность к инфаркту.
- B_1, B_2 Здоровые пациенты двух типов.

Решается задача оценки параметров функции регрессии и выбора признаков при многоклассовой классификации.

Для каждого пациента измеряются концентрации белков $K, L, N, \dots$, абсорбированных на поверхности кровяных телец:

		K	L	...	N
A_1	001	x_{11}	x_{12}	...	x_{1n}
A_1	002	x_{21}	x_{22}	...	x_{2n}
...	...	...	...	...	
A_3	i	x_{i1}	x_{i2}	...	x_{in}
A_3	$i+1$	$x_{(i+1)1}$	$x_{(i+1)2}$	...	$x_{(i+1)n}$
...	...	...	...	...	

Требуется, имея набор признаков $\mathbf{x}_i = [x_{i1}x_{i2}\dots x_{in}]$ определить, к какой из четырех групп A_1, A_3, B_1, B_2 относится пациент.

Чтобы перейти от задачи многоклассового прогнозирования к двуклассовой задаче, рассмотрим всевозможные пары групп:

$A_1\ A_3$	$A_3\ B_1$
$A_1\ B_1$	$A_3\ B_2$
$A_1\ B_2$	$B_1\ B_2$

Задача классификации решается для каждой пары групп.

Дана выборка $D = \{\mathbf{x}_i, y_i\}, i = 1, \ldots, m$, где
$y_i \in \{0, 1\}$ — объект,
$\mathbf{x}_i \in \mathbb{R}^n$ — описание объекта.
Разделение выборки на классы осуществляется с помощью логистической регрессии.
Предполагается, что $\mathbf{y} = [y_1, \ldots, y_m]^T$ — случайный вектор с независимыми компонентами $y_i \sim \mathcal{B}(\theta_i)$. Определим θ_i как

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

где $\mathbf{w}$ — вектор параметров логистической регрессии.

Алгоритм классификации имеет вид

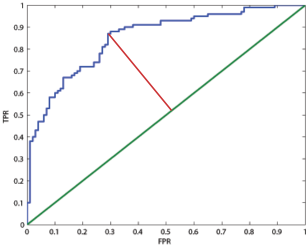
$$a(\mathbf{x}) = \operatorname{sign}\big(f(\mathbf{x}, \mathbf{w}) - c_0\big),$$

где c_0 — пороговое значение (cut-off) функции регрессии.

$\mathcal{I} = \{1, 2, \ldots, m\}$ — множество индексов объектов,
$\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$, $\mathcal{L}$ — обучающее подмножество, $\mathcal{T}$ — тестовое.

Параметры $\mathbf{w}$ оцениваются на подвыборке $D_{\mathcal{L}}$, а качество прогноза вычисляется на подвыборке $D_{\mathcal{T}}$.

Для оценки качества прогноза используется площадь AUC под кривой ROC (area under curve).



где
TPR — доля правильно классифицированных в пользу заданного класса объектов
FPR — доля ошибочно классифицированных в пользу данного класса объектов выборки.
Отрезок [(0,0),(1,1)] соответствует отказу от принятия решения в пользу какого-нибудь из классов.

Стоимость полного обследования одного пациента составляет 3400 евро. Уменьшение числа измеряемых признаков позволит:

1. Сократить расходы на обследование пациента.
2. Увеличить количество пациентов в выборке.

Выбор признаков осуществляется полным перебором. Оптимальный набор $\hat{\mathcal{A}}$ определяется как

$$\hat{\mathcal{A}} = \arg\max_{\mathcal{A} \subseteq \mathcal{I}} \mathrm{AUC}(\mathcal{A}) \text{ при условии } |\mathcal{A}| = \mathrm{const},$$

где $\mathcal{A} \subseteq \mathcal{J} = \{1, 2, \ldots, n\}$ — некоторое подмножество индексов признаков.

Для объекта $\mathbf{x}_{m+1}$ построим таблицу

$$\begin{array}{c|cccc} & A_1 & A_3 & B_1 & B_2 \\ \hline A_1 & - & 0 & 0 & 1 \\ A_3 & 1 & - & 1 & 1 \\ B_1 & 1 & 0 & - & 0 \\ B_2 & 0 & 0 & 1 & - \end{array}.$$

Присвоим классам A_1, A_3, B_1, B_2 номера $1, 2, 3, 4$. Тогда

$$\text{class}(\mathbf{x}_{m+1}) = \arg \max_{l \in \{1,\ldots,4\}} \sum_{k=1}^{4} a_{lk}(\mathbf{x}_{m+1}),$$

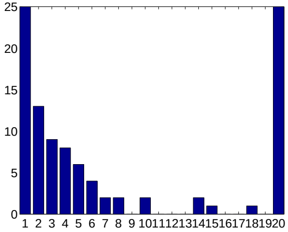
где $a_{lk}(\mathbf{x}) \in \{0, 1\}$, $l, k \in \{1, \ldots, 4\}$ — результат классификации между парой классов (l, k).

Вероятность можно оценивать, используя K оптимальных наборов для каждого класса.

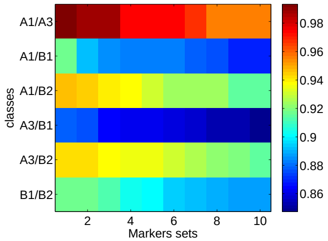
$$\begin{array}{c|cccc} & A_1 & A_3 & B_1 & B_2 \\ \hline A_1 & - & 0,1,0 & 0,0,0 & 1,1,0 \\ \dots & \dots & \dots & \dots & \dots \end{array}.$$

$$\downarrow$$

$$\begin{array}{c|cccc} & A_1 & A_3 & B_1 & B_2 \\ \hline A_1 & - & 0 & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \end{array}.$$



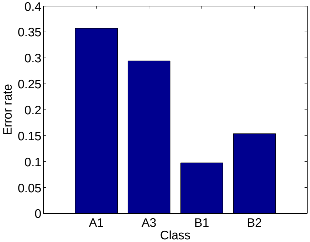
a.



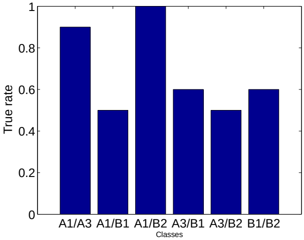
b.

a. Количество появлений каждого признака в $K = 5$ лучших наборов для какой-либо пары классов.

b. По горизонтали — наборы признаков, вошедшие в $K = 5$ лучших для какой-либо пары классов. Цветом обозначено значение AUC, соответствующее набору.



a.



b.

a. Количество неправильно классифицированных при многоклассовой классификации пациентов каждого класса.
b. Количество правильно классифицированных пациентов каджого класса при выполнении двуклассовой классификации между всевозможными парами.

Методы оценки объема выборки

1. Метод доверительных интервалов.
2. Метод оценки объема выборки в логистической регрессии.
3. Метод скользящего контроля.
4. Сравнение плотностей распределения параметров модели на различных подвыборках.

Пусть $D = \{(x_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$.

$$\bar{x} = \frac{1}{2} \sum_{i=1}^{m} x_i.$$

При известных μ и σ случайная величина имеет распределение

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

Получаем

$$m^* = \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2,$$

где $P\left\{|Z| \geq z_{\alpha/2}\right\} = \alpha$.

Предполагается, что

1. $x_i \sim \mathcal{N}(\mu, \sigma^2)$.
2. Значения μ и σ^2 известны.

Более вероятно, что

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

Фиксируем множество индексов признаков $\mathcal{A}$ и индекс $j \notin \mathcal{A}$.

$$H_0 : w_j = 0 \ \rightarrow \mathbf{w}_{\mathcal{A}}, \ H_1 : w_j \neq 0 \rightarrow \mathbf{w}_{\mathcal{A}^*}, \text{ где } \mathcal{A}^* = \mathcal{A} \bigcup \{j\}.$$

Оценим вектор $\boldsymbol{\theta}$ параметров бернуллиевского распределения:

$$H_0 : \boldsymbol{\theta} = f(X_{\mathcal{A}}^T \mathbf{w}_{\mathcal{A}}), \ H_1 : \boldsymbol{\theta} = f(X_{\mathcal{A}^*}^T \mathbf{w}_{\mathcal{A}^*}).$$

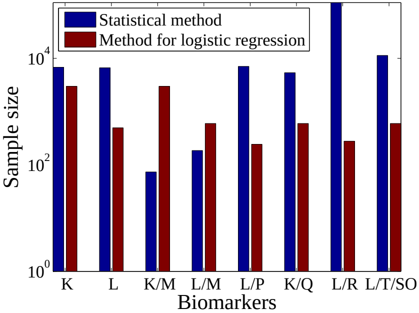
Важно лишь пороговое значение функции регрессии c_0, поэтому выберем в качестве тестовой статистики

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 c_0 / m}}, \ \hat{p} = \frac{1}{m} \sum_{i=1}^{m} y_i$$

где $\hat{p}$ — ОМП для параметра p, $c_0 = 1 - p_0$ — пороговое значение функции регрессии.
Тогда

$$m^* = \frac{p_0 c_0 \left(z_{1-\alpha/2} + z_{1-\beta} \sqrt{\frac{p_1 c_1}{p_0 c_0}} \right)^2}{(p_1 - p_0)^2}.$$

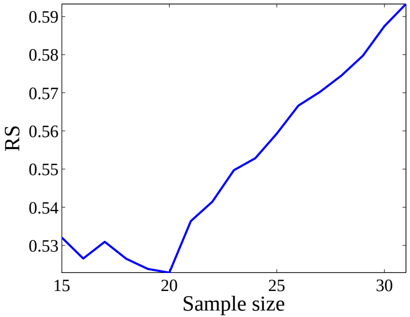
Результаты метода доверительных интервалов и метода для логистической регрессии



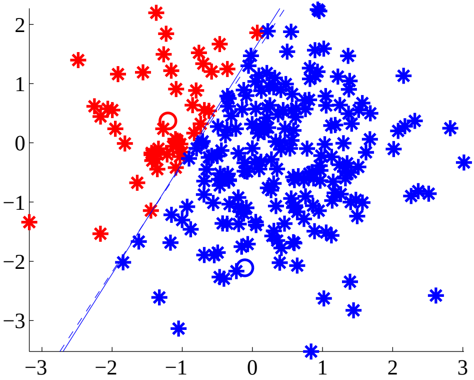
Метод доверительных интервалов: $m^* \sim 10^5$.
Метод для логистической регрессии: $m^* \sim 10^3$.

Введем величину

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



Вывод: $m^* \geq 30$.



В выборку были добавлены два новых объекта.
Положение разделяющей гиперплоскости определяется выражением

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

Для оценки сходства функций плотности $p_1(\mathbf{w})$ и $p_2(\mathbf{w})$ воспользуемся расстоянием Куллльбака-Лейблера

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

Плотность вероятности появления данных есть

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

Пусть также $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2 I_{|\mathcal{A}|})$, и его плотность имеет вид

$$p(\mathbf{w}|f_{\mathcal{A}}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|\mathcal{A}|}{2}} \exp(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T \alpha I(\mathbf{w}-\mathbf{w}_0)),$$

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

Для нахождения $p(\mathbf{w}|D, \alpha, f_{\mathcal{A}})$, воспользуемся формулой Байеса

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

где $p(D|\mathbf{w}, f_{\mathcal{A}})$ —правдоподобие данных,
$p(\mathbf{w}|\alpha, f_{\mathcal{A}})$ —вероятностные плотности параметров модели,

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

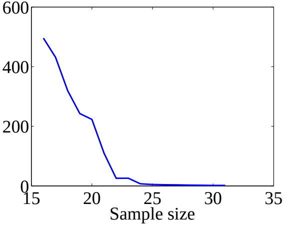
Тогда

$$p(\mathbf{w}|D, f_{\mathcal{A}}) = \frac{p(y|\mathbf{x}, \mathbf{w}, f_{\mathcal{A}})p(\mathbf{w}|f_{\mathcal{A}}, \alpha)}{Z(\alpha)} =$$

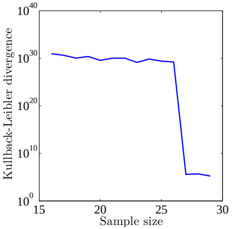
$$= \frac{\alpha^{\frac{|\mathcal{A}|}{2}}}{(2\pi)^{\frac{|\mathcal{A}|}{2}} Z(\alpha)} \exp(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \alpha I(\mathbf{w} - \mathbf{w}_0)) \prod_{i=1}^{m} \theta_i^{y_i}(1 - \theta_i)^{1-y_i},$$

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

Оценка объема выборки с помощью расстояния Кулльбака-Лейблера



a.



b.

a. Усредненное еклидово расстояние между параметрами модели

$$||\mathbf{w}_m - \mathbf{w}_{m+1}|| = \sqrt{\sum_{i=1}^{|\mathcal{A}|} \left(w_i^m - w_i^{m+1}\right)^2},$$

b. Усредненное расстояние Кулльбака-Лейблера.

Вывод: $m^* \geq 30$.

Результаты

1. Предложен алгоритм выбора признаков для многоклассовой классификации
2. Найдены оптимальные наборы признаков для выборки пациентов с инфарктом миокарда
   Публикации:
   1. Мотренко А.П., Многоклассовый прогноз вероятности наступления инфаркта и оценка объема выборки // JMLDA, декабрь 2011
   2. Мотренко А.П., Стрижов В.В., Многоклассовый прогноз вероятности наступления инфаркта // Известия Тульского государственного университета, март 2012
3. Предложен метод оценки объема выборки, основанный на исследовании пространства параметров модели.
   Публикация:
   Мотренко А.П. Оценка необходимого объема выборки пациентов при прогнозировании сердечно-сосудистых заболеваний // JMLDA, май 2012.