Оценка оптимального объёма выборки для задач классификации*

***Харатян А. С., Катруца А. М.*[1], *Стрижов В. В.*[2]**
haratyan.as@phystech.edu; aleksandr.katrutsa@phystech.edu;
strijov@phystech.edu
[1]Московский физико-технический институт, Москва, Россия;[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, Москва, Россия

В статье рассматривается задача выбора оптимального числа объектов выборки для их классификации. Исследуется использование порождающих и разделяющих вероятностных моделей бинарной классификации. Обсуждается проблема медицинской диагностики пациентов. Определяется понятие достаточности объёма выборки. Показывается, какими методами возможно выбрать оптимальное количество объектов, обеспечивающее необходимую точность классификации объектов . В работе рассматривается, применение каких критериев выявляет наилучшее качество классификации. Приводится теоретическое и практическое обоснование предложенных критериев. Используется модель логистической регрессии.

Ключевые слова: *определение оптимального объёма выборки, логистическая регрессия, расстояние Кульбака-Лейблера.*

1 Введение

Работа посвящена оценке оптимального объёма исследуемой выборки применительно к проблемам медицинской диагностики. Рассматриваются биомедицинские данные пациентов как выборка. Каждый пациент обладает набором признаков. Получение данных о пациентах требует немалых средств. В случае, если количество данных избыточно, то их измерения приносят крайне неоправданные расходы. В связи с этим поднимается вопрос оптимального количества измерений. Ввиду дороговизны анализов всех признаков оценка измерений должна быть точной.

Для нахождения оценки используется модель логистической регрессии [1]. Стандартной практикой является использование статистических методов [2] для оценивания объёма данных при помощи логистической регрессии. Введём понятие устойчивости модели в отношении объёма выборки. Будем называть модель устойчивой, если при изменении малом изменении объёма параметры модели меняются незначительно. Если размер выборки крайне мал и недостаточен, то параметры модели меняются скачкообразно при увеличении объёма. Соответственно, увеличивая объём выборки, мы повышаем устойчивость модели. В качестве показателя устойчивости для моделей будем использовать расстояние Кульбака-Лейблера. Для того чтобы показать отличие моделей в устойчивости будем использовать разность усредненных значений расстояний К-Л, вычисленных на разных выборках одного и того же объёма. Если объекты порождены одинаковым распределением, то при росте объёма выборок разность расстояний К-Л между моделями падает. Достигнув необходимого показателя устойчивости, можно легко вычислить оптимальный размер данных.

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Консультант: Катруца А. М.

Для вычислительного эксперимента используются реальные данные 569 пациентов с 30 признаками и метками об опухоли молочной железы: доброкачественная или злокачественная. Как описано выше, будем обучать модели на разных подвыборках и после достижения необходимого уровня устойчивости получим оптимальный объём данных.

2 Постановка задачи классификации

Пусть у нас задана выборка $D = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, m\}$ с объёмом m объектов(пациентов), каждый из которых описывается n признаками, $\mathbf{x}_i \in \mathbb{R}^n$ и принадлежит одному из двух классов: $y_i \in \{0, 1\}$. Модель логистической регрессии предполагает, что вектор целевой переменной $\mathbf{y} = [y_1, \ldots, y_m]^T$ имеет распределение Бернулли, $y_i \sim \mathscr{B}(\theta_i)$ с плотностью распределения

$$p(y|\boldsymbol{\omega}) = \prod_{i=1}^{m} \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \tag{1}$$

Плотность вероятности зависит от вектора параметров ω. Зная ω, можно вычислить вероятность принадлежности к классу

$$\theta_i = f\left(\mathbf{x}_i^T \omega\right) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^T \omega\right)} \tag{2}$$

Пользуясь принципом максимального правдоподобия, можем вычислить функцию ошибки для уравнения (1)

$$E(\boldsymbol{\omega}) = -\ln p(\mathbf{y}|\boldsymbol{\omega}) = -\sum_{i=1}^{m} \left(y_i \ln \theta_i + (1 - y_i) \ln (1 - \theta_i)\right) \tag{3}$$

Чтобы найти вектор параметров логистической регрессии $\widehat{\omega}$, необходимо решить оптимизационную задачу:

$$\hat{\omega} = \arg\min_{\omega \in \mathbb{R}^n} E(\omega) \tag{4}$$

Тогда алгоритм классификации определяется следующим образом:

$$a(\mathbf{x}, c_0) = sign\left(f(\mathbf{x}, \omega) - c_0\right) \tag{5}$$

где c_0 - пороговое значение функции активации [3] .

3 Постановка задачи определения оптимального объёма выборки

Допустим, что задано m объектов выборки D. Количество этих объектов недостаточно для обучения устойчивой модели логистической регрессии. Необходимо, имея m объектов, найти такое число m^*, что показатель устойчивости модели при увеличении числа m^* меняется незначительно. Задача нахождения данного числа объектов представляется следующим образом:

$$m^* = \min m \in \mathbb{N} \;\; \forall k \in \mathbb{N} \to \rho(\mathrm{p}(\boldsymbol{\omega}|m^*), \mathrm{p}(\boldsymbol{\omega}|m^* + k)) < \varepsilon \tag{6}$$

где ρ - некоторая функция расстояния между распределениями, ε - заранее заданный порог устойчивости, $\mathrm{p}(\boldsymbol{\omega}|m^*)$ - распределение весов модели, обученной на выборке размером m^*.

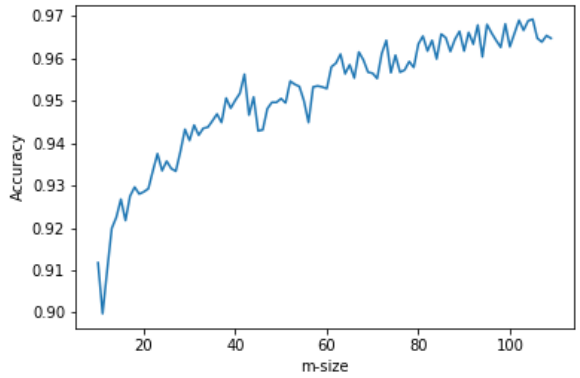
4 Вычислительный эксперимент

В качестве функции расстояния между распределениями будем использовать дивергенцию Кульбака-Лейблера. Предполагается, что веса модели ω порождаются многомерным нормальным распределением. Учитывая это, можно легко посчитать расстояние между двумя распределениями параметров со средними значениями $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ и ковариационными матрицами A_1, A_2 :

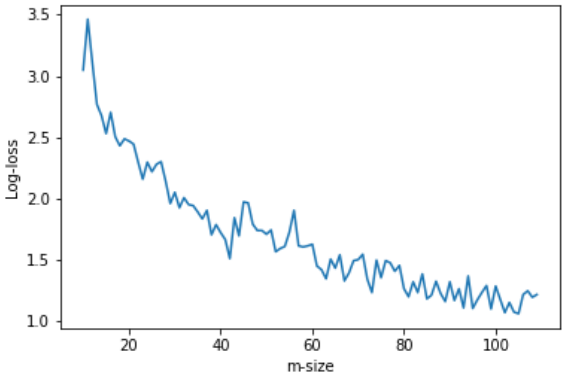
$$\rho(\mathsf{p}(\boldsymbol{\omega}|m), \mathsf{p}(\boldsymbol{\omega}|m+k)) = KL(\mathcal{N}(\hat{\boldsymbol{\omega}}_1, \boldsymbol{A}_1)||\mathcal{N}(\hat{\boldsymbol{\omega}}_2, \boldsymbol{A}_2)) =$$

$$\frac{1}{2}\left(tr(\boldsymbol{A}_2^{-1}\boldsymbol{A}_1) + (\hat{\boldsymbol{\omega}}_1 - \hat{\boldsymbol{\omega}}_2)^T\boldsymbol{A}_2^{-1}(\hat{\boldsymbol{\omega}}_1 - \hat{\boldsymbol{\omega}}_2) - n + \ln\left(\frac{\det(\boldsymbol{A}_2)}{\det(\boldsymbol{A}_1)}\right)\right)$$

В эксперименте применяются данные о 569 пациентах с метками о произрастании опухоли молочной железы. Стоит отметить, что изначально мы рассматриваем задачи медицинской диагностики, где данных может быть намного меньше. Однако в данном случае мы воспользуемся большим объёмом, поскольку нам требуется точнее оценить среднее качество классификации, достигаемое на фиксированной длине выборке. Для этого мы будем много раз обучать модель регрессии, выбирая случайным образом определенное количество объектов, начиная с 10 и заканчивая 100. Для каждой длины выборки будем отмечать точность классификации, полученную на других случайных 100 объектах из 569, логистическую функцию потерь и расстояние Кульбака-Лейблера. Результаты эксперимента и графики зависимости от объёма выборки представлены ниже.



(a) Точность классификации при разных объёмах

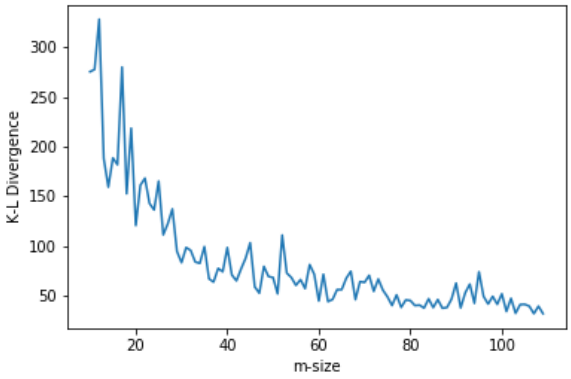


($б$) Функция ошибки log-loss

Нетрудно увидеть, что с увеличением объёма выборки качество классификации повышается, однако при добавлении новых данных качество предсказани и ошибки изменяются медленнее. Например, на графике точности количества объектов порядка 20-25 хватает для достижения почти 93% точности. При дальнейшем пополнении выборки, когда объём увеличивается в 5 раз, точность выше только на 3%. Ввиду такого факта следует, что не имеет смысла брать выборки очень большого масштаба для решения реальных задач прогнозирования. В целом, обычно 95% точности является очень хорошим результатом классификации пациентов в медицинской диагностике. Точность выше обычно требуется на data-science соревнованиях, где каждая доля процента крайне важна. Заметим, что в модели логистической регрессии используется стохастический градиентный спук, который чувствителен к параметрам модели. Чтобы процесс nt оказался парализованным и быст-

рее сходился за определенное число шагов, была произведена нормализация признаков. Из каждого признака вычитается среднее по признаку всех объектов выборки и делится на стандартное отклонение.

Достижение полученных результатов точности на выборках малого объёма появляется не случайно. Этим обоснован выбор модели для нашего эксперимента. Логистическая регрессия очень хорошо работает при выборках небольшого объёма. Другие модели, как например, дерево решений, могут требовать в тысячи раз больше данных. Такой исход можно увидеть в работе [4]. Однако не всегда при помощи простой логистической регрессии удаётся добиться крайне завышенных результатов, как например 99%. Для этой цели уже следует использовать усовершенствованные методы классификации. В качестве примера таких методов можно привести многослойный перцептрон.



(a) Расстояние Кульбака-Лейблера.

Стоит внимательно посмотреть на график расстояния Кульбака-Лейблера. Из рисунка можно увидеть, что примерно после объёма 40 метрика начинает меняться значительно слабее. Таким образом, мы получили оптимальный номер объектов для нашей выборки в модели логистической регрессии.

5 Выводы

В работе описан алгоритм оценивания оптимального объёма выборки для исследований в медицине. В эксперименте были рассмотрены реальные данные, при помощи которых удалось выявить оптимальное количество элементов для обучения логистической регрессии(приблизительно 40 объектов)

Литература

[1] *Hosmer Jr David W, Lemeshow Stanley, Sturdivant Rodney X.* Applied logistic regression. — 2013. Vol. 398.

[2] *Demidenko Eugene.* Sample size determination for logistic regression revisited // Statistics in medicine, 2007. Vol. 26. No. 18. P. 3385–3397.

[3] *Motrenko Anastasiya, Strijov Vadim, Weber Gerhard-Wilhelm.* Sample size determination for logistic regression // Journal of Computational and Applied Mathematics, 2014. Vol. 255. P. 743–752.

[4] *Sug Hyontai.* An effective sampling method for decision trees considering comprehensibility and accuracy // W. Trans. on Comp., 2009. Vol. 8. No. 4. P. 631–640. URL: `http://dl.acm.org/citation.cfm?id=1558756.1558762`.

Received