Раннее прогнозирование достаточного объема выборки для обобщенно линейной модели

Жолобов В. А. Малиновский Г. С Стрижов В. В.

Исследуется проблема планирования эксперимента. Задача ранего прогнозирования важна в медицинском применении, особенно в случаях дорогостоящих измерений иммунных биомаркеров. Решается задача оценивания достаточного объема выборки для поиска адекватной регрессионной модели. Предполагается, что выборка является простой. Она описывается адеватной моделью. Иначе, выборка порождается фиксированной вероятностной моделью из известного класса моделей. Объем выборки считается достаточным, если модель восстанавливается с достаточной достоверностью. Исследуется зависимость функции ошибки от объема данных. Исследуется зависимость модели от редуцированной матрицы ковариации параметров GLM. Требуется, зная модель, оценить достаточный объем выборки на ранних этапах сбора данных. Предложены алгоритмы оценки достаточного объема выборки. Проведен вычислительный эксперимент с использованием синтетических данных.

Ключевые слова: *sample size forecasting; empirical sample size, active learning; sample size*

1 Введение

Работа посвящена задаче оценивания достаточного объема выборки на раннем этапе сбора данных. Задача возникла из условия, когда необходимо провести крупное исследование, а сбор данных является дорогостоящим. Для примера можно взять медицинское исследование, такой как анализ крови. Существуют такие виды анализа крови, которые стоят достаточно приличных денег для людей. Для того, чтобы снизить стоимость данных для исследований в несколько раз необходимо построить модель, а для модели нужно собрать выборку. В данной работе рассматривается задача построения алгоритма для предсказания оптимального набора данных при заданной модели. Предлагаемый в данной работе метод должен на малой выборке спрогнозировать ошибку на пополняемой большой. Выборка считается простой, то есть удовлетворяет простому распределению. Предлагается использовать два разных метода: полного перебора и генетический алгоритм.

Ранее задача прогнозирования достаточного объема выборки решалась В работе [?] предложен метод, основанный на технике кросс-валидации и расстоянии Кульбака-Лейблера между двумя распределениями параметров модели, оцениваемых на аналогичных подмножествах данных. Задача информационного поиска решалась в работах [?, ?]. Для создания простых структурированных функций информационного поиска используется модернизированный генетический алгоритм. Эвристика генетического алгоритма заключается в том, что он способен работать при стагнации признаков.

В данной работе используются два метода для аппроксимации эмпирической функции ошибки. Основной из них — это метод полного перебора. Необходимо подобрать такую функцию, которая являетсямонотонной и достаточно гладкой, то есть гарантируется непрерывная дифференцируемость до второго порядка. Метод заключается в том, что он аппроксимирует зависимость функции ошибки от объема данных по малому объему выборки, чтобы с достаточной точностью предсказывать ее поведение. Считается, что модель в этой задаче задана и зависит от редуцированной матрицы ковариации параметров GLM. Также предложен способ генерации такой функции через генетический алгоритм.

Вычислительный эксперимент проводится на синтетических данных *Boston Housing* и *Diabets*. Вначале реализуем метод полного перебора. Разделяем выборку на два множества. Строим два графика поверхности выборок: первую получаем с помощью бутстрепа [?] для подвыборки фиксированного объема, вторую через аппроксимацию. Чтобы получить аппроксимирующую поверхность, решается оптимизационную задачу. Затем повторяем действия, используя уже для поиска аппроксимирующей функции генетический алгоритм. Решение этой задачи позволит находить оптимальное значение объема выборки.

2 Постановка задачи раннего прогнозирования достаточного объема выборки для обобщенно линейной модели

Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, являющаяся простой, *i.i.d.*, семплированная из экспоненциального распределения с неизвестными параметрами. Обозначим объем выборки как $m_{\mathfrak{D}} = |\mathfrak{D}|$. Решается задача раннего прогнозирования минимально необходимого объема выборки. Под ранним прогнозированием понимается прогнозирование при однократно заданном объеме m_0 такого объема m^*, который необходим для построения адекватной модели. Задана модель $f(\mathbf{w}, x)$ с параметрами из распределения $P(\mathbf{w}|f, \mathfrak{D}, m)$ и функция регрессии $f : \mathbf{x} \mapsto y$:

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon_i$$

$f : w, \mathbf{x} \mapsto y$. Для определения параметра $\mathbf{w}$ задаются функции ошибки $S(\mathbf{w})$ точности аппроксимации в задачах линейной и логистической регрессиях:

$$S = S(\mathfrak{D}, f, w^*, m). \tag{1}$$

Требуется найти параметры регрессионной модели $\mathbf{w}^*$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(w|\mathfrak{D}, f), \tag{2}$$

где S — теоретическая функция ошибки.

Рассматривается две обобщенных линейных модели: линейная регрессия и логистическая регрессия. Пусть дано множество из m пар (x_i, y_i), $i = 1, \ldots, m$, а также пусть назначена линейная модель с аддитивной случайной величиной ε_i. Требуется найти параметры регрессионной модели $\mathbf{w}^*$. Ошибка в случае линейной регрессии

$$S(\mathbf{w}) = ||\mathbf{f} - y||_2^2. \tag{3}$$

Постановка задачи логистической регрессии. Для обучения линейного классификатора решается задача минимизация эмпирического риска с функцией потерь в двухклассовом случае

$$S(\mathbf{w}) = y \ln(f) + (1 - y) \ln(1 - f) \tag{4}$$

В общем случае функция потерь

$$S(\mathbf{w}) = \sum_{i,k} y_{i,k} \ln f_{i,k} \tag{5}$$

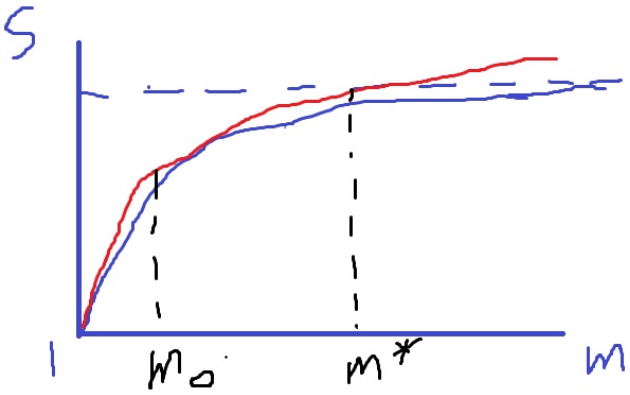
В задаче поиска оптимального состава признаков $\mathbf{x}_{\mathcal{A}} = [x_{1\mathcal{A}} \ldots x_{n\mathcal{A}}]^T$ требуется оптимизировать набор признаков $n^* = |\mathcal{A}|$ и

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subset\{1,\ldots,n\}} Q(\mathcal{A}|\mathbf{w}, \mathfrak{D}). \tag{6}$$

Обозначим зависимость значения функции ошибки (1) от объема выборки как

$$S(m) = S(m|\mathbf{w}, \mathcal{A}, \mathfrak{D})$$

Так как значение функции S зависит от объема выборки, то и при каждом значении m решается отдельная обобщенно-линейная задача (2). Построим график эмпирической зависимости $\mathsf{E}S$, $\mathsf{D}S$ от значения объема выборки m при заданном оптимальном $\hat{w}$ функция ошибки является эмпирической стабилизирующейся функцией ошибки. Предполагается, что эмпирическая зависимость имеет вид



Определим понятие стабилизации функции ошибки S согласно

Определение 1. Стабилизирующаяся функция ошибки S — *это функция, обладающая таким свойством*

$$\lim_{m \to +\infty} S(m) = r, \tag{7}$$

где r — заданное значение горизонтальной асимптоты.

Определение 2. Функции ошибки S сходится к r — значение $m^* < +\infty$, начиная с которого выполняется

$$\frac{dS}{dm} = 0, \quad \forall m \geqslant m^*, \tag{8}$$

а также выполняется условие предела

$$\forall \varepsilon > 0 \quad \exists m' > 1 : \quad ||S(m) - r||_1 < \varepsilon, \quad \forall m \geqslant m'. \tag{9}$$

В задаче раннего прогнозирования предполагается заданной ожидаемая точность $r \leqslant S$. По заданной выборке $\mathfrak{D}(m_0)$ необходимо построить функцию $\varphi(m)$, аппроксимирующую эмпирическую функцию $S(m)$. Функция $\varphi : \mathbb{N} \to \mathbb{R}_+$ выбирается из заданного класса E с ограничением, связанным с принадлежностью функции $\varphi(m)$ классу дважды непрерывно дифференцируемых функций $E \subset \mathbb{C}^2(\mathbb{R}_+)$ Для поиска функции решается условная оптимизационная задача

$$\varphi^*(m) = \underset{\varphi(m) \in E, m \in [1, m_0]}{\arg\min} ||\varphi(m) - S(m)||_1. \tag{10}$$

Для решения задачи выбирается функция $\varphi^*(m)$ и ищется прообраз в значении заданной точности

$$m^* = (\varphi^*)^{-1}(r), \tag{11}$$

который и является прогнозом минимального объема.

3 Описание алгоритма

При решении задачи раннего прогнозирования достаточного объема выборки в условиях неизвестной структуры состава признаков $\mathcal{A}$ модели f вводится отношение частичного порядка на множестве признаков. Из имеющегося набора выборок выделяется методом бутстрепа подвыборка D_m заданного объема m. При поиске оптимальной функции φ^* необходимо учитывать конфигурацию признаков.

Определение 3. *Бутстреп* — метод исследования вероятностных распределений, основанный на многократном семплирования выборок равномощно равномерным методом на базе имеющейся выборки.

Определение 4. *Бутстреп* — получение выборки равномощным равномерным методом семплирования.

Элементы подвыборки $i \sim U[1, \dots, m]$ образуют подвыборку мощностью D. Здесь корректно делать бутстреп в силу простоты выборки и порождения ее единой моделью (i.i.d.).

Определение 5. *Отношение частичного порядка* (отношение нестрогого частичного порядка) — бинарное отношение R на множестве X, удовлетворяющее условиям

1. Рефлексивность: $\forall x \in X : xRx$
2. Антисимметричность: $\forall x, y \in X : xRy \wedge yRx \Rightarrow x = y$
3. Транзитивность: $\forall x, y, z \in X : xRy \wedge yRz \Rightarrow xRz$

Необходимость задания отношения частичного порядка обусловлена присутствием коллериющих признаков. Задается отношение частичного порядка с помощью алгоритма Лассо. Он заключается в добавлении ограничения на регрессионные коэффициенты $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_n)$

$$\sum_{j=1}^{n} |\hat{w}_j| \leqslant \lambda, \tag{12}$$

где $\lambda \geqslant 0$ — параметр регуляризации. Чем меньше λ, тем большее число признаков $\hat{w}_i$ принимают нулевое значение. Так Lasso производит отбор информативных признаков.

В эксперименте предполагается присутствие трех видов функций ошибки. Определим каждую из них

Определение 6. *Эмпирическая функция ошибки* S — это реализация случайной величины (1) с математическим ожиданием $\mathsf{E}S$ и дисперсией $\mathcal{D}S$ при заданной выборке.

Определение 7. *Теоретическая функция ошибки* S — это дважды дифференцируемая по m функция при заданном распределении $p(y)$

Определение 8. *Аппроксимация функции ошибки* $\varphi(m)^*$ — это функция из семейства E.

Математическое ожидание функции ошибки $\mathbb{E}S$ задается в виде интеграла

$$\mathsf{E}S = \int_{\mathbb{R}} sp(s)ds. \tag{13}$$

В случае, когда неизвестна функция распределения $p(s)$, используется приближение математического ожидания. Используется бутстреп для получения k подвыборок размеров m и в результате математическое ожидание

$$\mathsf{E}S(m) = \frac{1}{k}\sum_{i=1}^{k} S_i. \tag{14}$$

Такое допущение возможно в связи семплирования выборки. Таким образом вычисляется значение эмпирической функции ошибки S. Дисперсия эмпирической функции ошибки S

$$\mathsf{D}S(m) = \frac{1}{k}\sum_{i=1}^{k}(S_i - \mathsf{E}S(m))^2. \tag{15}$$

Описания алгоритма

1. Вычисление эмпирической функции ошибки в виде (13) и (15).
2. Подбор параметров семейства функций (16). Решение оптимизационной задачи (17).
3. Поиск искомого объема m^* на графике функций ошибок (11).

Для аппроксимации эмпирической функции S строится параметрическое семейство E вида

$$E = \{\exp(w_1 + w_2 \ln m) + w_3 + w_4 \ln m | w_2 > 0\} \tag{16}$$

Поиск функции $\varphi(m)^*$ осуществляется с помощью поиска оптимального набора параметров $\mathbf{w} = (w_1, w_2, w_3, w_4)$. Формально задача оптимизации ставится в виде

$$\min_{\mathbf{w}} ||\exp(w_1 + w_2 \ln m) + w_3 + w_4 \ln m - \mathsf{E}S(m)||_1$$

$$\text{s.t. } w_2 > 0 \tag{17}$$

Решением задачи оптимизации будут параметры для искомой аппроксимации эмпирической функции $\varphi(m)^*$. По найденным эмпирической функции ошибки $\mathbb{E}S$ и ее аппроксимации $\varphi(m)^*$ строим график зависимости этих функций от m. В точке m^*, где функции будут пересекаться, и будет ответ на задачу раннего прогнозирования достаточного объема выборки для обобщенной линейной модели. По расположению m^* относительно m можно судить о заданном объеме выборки

1. Если $m^* < m$, то выборка избыточная;
2. Если $m^* > m$, то выборка имеет достаточный объем;
3. Если точки пересечения нет, то у выборки нет необходимого объема для предсказания;

4 Вычислительный эксперимент

Эксперимент проводится в несколько этапов. На первом этапе строится линейная регрессия (2) по выборке из Таблицы 1. Полученная функция ошибки модели используется

в дальнейшем. На втором этапе фиксируется выборка $\mathfrak{D}$ и построенная модель f. Затем строится математическое ожидание эмпирической функции ошибки $\mathbb{E}S$ и ее дисперсия $\mathcal{D}S$. Так проводится проверка гипотезы о стабилизации эмпирической функции ошибки S. На третьем этапе аппроксимируется эмпирическая функция ошибки (1) на заданном семействе аппроксимирующих функций для значений объема выборки $m < m^*$. На четвертом этапе решается задача (10) с учетом решения вспомогательной задачи (6).

Для анализа алгоритма предлагается использовать три различных выборки, описания которых приведены в Таблице 1

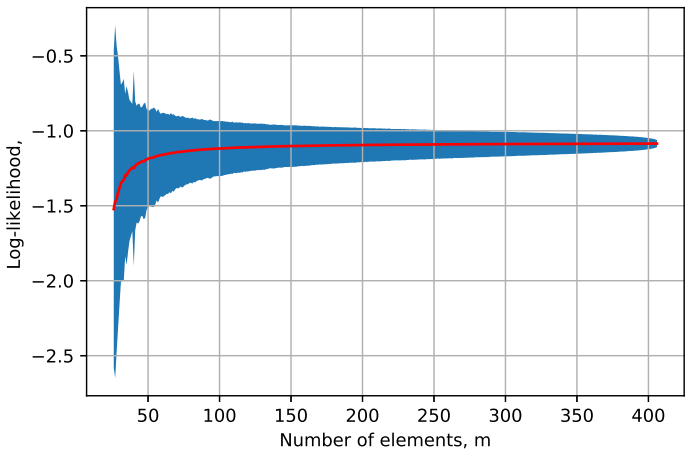
Таблица 1 Описание выборок

Выборка	Тип задачи	Размер выборки	Число признаков
Boston	Регрессия	506	13
Diabets	Регрессия	167	4
Synthetic 1	Регрессия	50000	4

Синтетическая выборка Synthetic 1 генерирует данные из равномерного распределения на интервале $[0, 1)$.

4.1 Случай заданного семейства аппроксимирующих функций

Для начала фиксируется выборка Boston и строится зависимость эмпирической ошибки линейной регрессии от объема выборки с учетом дисперсии.

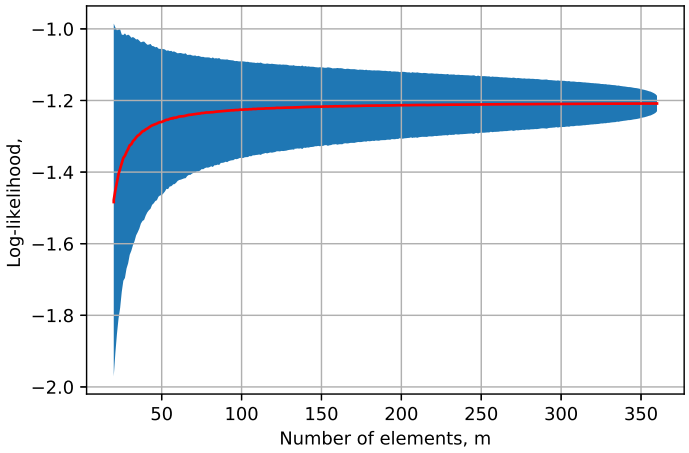


Видно, что гипотеза о стабилизирующейся эмпирической функции ошибки подтверждается экспериментально на трёх выборках. Модель, аппроксимирующая эмпирическую функцию ошибки задается таким образом

$$E = \{\exp(w_1 + w_2 \ln m) + w_3 + w_4 \ln m | w_2 > 0\} \tag{18}$$

При заданном объеме выборки m_0 производим подбор набора коэффициентов $\mathbf{w} = = (w_1, w_2, w_3, w_4)$ Так ищется функция $\varphi(m)$ из задачи (10) Затем по заданной точности r восстанавливаем прообраз функции $\varphi(m)$. Найденное значение и есть ответ на задачу. Найденную функцию $\varphi(m)$ построим вместе с дисперсией и эмпирической функцией ошибки.

Для выборки Diabets строится зависимость эмпирической ошибки линейной регрессии от объема выборки с учетом дисперсии.



Ищется аппрокцимация из параметрического семейства E. Полученное значение строится на графике.