Раннее прогнозирование достаточного объема выборки для обобщенно линейной модели

Жолобов В. А. Малиновский Г. 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. Предлагаются алгоритмы оценки достаточного объема выборки. Проведен вычислительный эксперимент с использованием синтетических данных.

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. Поэтому 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.

Кроме этих методов ранее задача прогнозирования достаточного объема выборкы решалась в работе [?]. Здесь был предложен метод, основанный на технике кросс-валидации и расстоянии Кульбака-Лейблера между двумя распределениями параметров модели, оцениваемых на аналогичных подмножествах данных. Похожая задача информационного поиска решалась в работах [?, ?]. Здесь 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. 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.

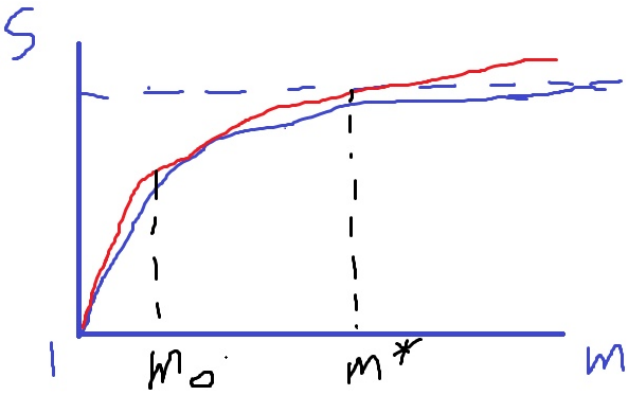
2 Постановка задачи

Задана выборка $\mathfrak{D} = [\mathbf{X}, \mathbf{y}] = \{(\mathbf{x}_i, y_i)\}$, являющаяся $i.i.d.$ произвольных значений, сгенерированных неизвестным распределением. Обозначим объем выборки как $m = |\mathfrak{D}|$. Решается задача раннего прогнозирования объема минимально необходимой выборки. Под ранним прогнозированием понимается прогнозирование при однократно заданном объеме m_0 объема m^*, необходимого для построения адекватной модели. При решении задачи прогнозирования оптимального объема выборки считаем заданным модель, то есть функцию $f : \mathbf{x} \mapsto y$. Для отображения f задаются функции ошибки, указывающая на точность модели в задачах линейной и логистической регрессиий

$$S = S(\mathfrak{D}, f, w^*, m) = ||f - y||^2$$

$$S(\mathbf{w}, \mathfrak{D}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}, \mathbf{x}_i)))$$

Предполагается фиксированность всех параметров, кроме объема данных. Предполагается, что функция ошибки является стабилизирующей эмпирической функцией. На графике такая функция продемонстрирована.



Рассматривается несколько постановок задач обобщенной линейной модели. Первой будет задача линейной регрессии. 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

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

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

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

где S — функция ошибки.

Постановка задачи логистической регрессии. Считаем, что здесь принята модель логистической регрессии, согласно которой свободные переменные x и зависимая переменная y связаны зависимостью

$$z = w_0 + \sum_{j=1}^{n} w_j x_j.$$

$$y = \frac{1}{1 + \exp(-z)} + \varepsilon,$$

Обозначим $p_i = f(w, x_i)$. Функция ошибки

$$S(\mathbf{w}, \mathfrak{D}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}, \mathbf{x}_i)))$$

В задаче поиска оптимального состава признаков $X_{\mathcal{A}}$ требуется оптимизировать набор признаков $n^* = |A|$ и

$$A^* = \arg\min Q(A|\mathbf{w}, \mathfrak{D})$$

Обозначим зависимость значения функции ошибки от выборки как $S(m) = S(m|\mathbf{w}, \mathcal{A}, \mathfrak{D})$. Так как функция S зависит от объема данных, то и при каждом значении m решается отдельная обобщенно-линейная задача $w^* = w^*(m)$.

В задаче раннего прогнозирования предполагается заданной ожидаемая точность r. 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 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

$$\varphi^*(m) = \underset{\varphi(m) \in E, m \in [0, m_0]}{\arg\min} ||\varphi(m) - S(m)||_1 \tag{1}$$

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

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

3 Вычислительный эксперимент

Эксперимент предлагается проводить в несколько этапов. На первом этапе строится линейная модель, в данном случае линейная регрессия. Полученная функция ошибки модели используется в дальнейшем. Второй этап состоит из аппроксимации функции ошибки предыдущего этапа при фиксированном числе параметров выборки на заданном семействе аппроксимирующих функций. Третий этап подразумевает аппроксимацию функции ошибки путем генетического алгоритма. На четвертом этапе решается задача с учетом решения вспомогательной задачи отбора важных признаков. На пятом этапе исследуется зависимость оптимального объема выборки от доступного размера выборки.

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

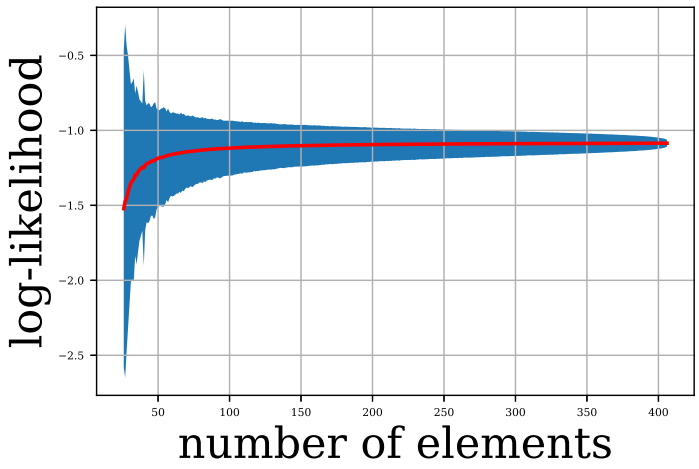
Таблица 1 Описание выборок

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

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

3.1 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 работает: подтверждается экспериментально на конкретной выборке. Модель аппроксимирующих функций задается таким образом

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

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 (1) Затем по заданной точности r восстанавливаем прообраз функции $\varphi(m)$. Найденное значение и есть ответ на задачу. В случае выборки Boston построим ее поверхность

В первом случае предполагаем число параметров фиксированным. В этом случае строим график зависимости функции ошибки от объема выборки на плоскости. Вертикальной штриховкой $---$ обозначается положение m_0. Есть наиболее используемый класс функций для оценки роста монотонных моделей. Если пытаться использовать каждую в отдельности, то итоговая ошибка будет слишком велика. Это видно на графике, как только объем выборки становится больше, чем m_0. В случае использовании их как в параметрическом семействе, то ошибка становится значительно меньше.