Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели.

Бучнев В. С., Грабовой А. В., Гадаев Т. Т., Стрижов В. В.

Исследуется проблема снижения затрат на сбор данных, необходимых для построения адекватной модели. Рассматриваются задачи обощенной линейной модели. Для решения этих задач требуется, чтобы выборка содержала необходимое число объектов. Требуется предложить метод вычисления оптимального обьема данных, соблюдая при этом баланс между точностью модели и и трудозатратами при сборе данных. Предпочтительны те методы оценки объемы, которые позволяют строить адекватные модели по выборкам возможно меньшего объема.

Ключевые слова: *Обобщенная линейная модель, размер выборки.*

1 Введение

При планировании эксперимента требуется оценить минимальный объём выборки — число производимых измерений набора показателей или признаков, необходимый для построениие сформулированных условий.

Существует большое количество оценки размера выборки. Например, тест множителей Лагранжа, тест отношения правдоподобия и тест Вальда. В работах [1–3] на основе данных методов построена оценка оптимального размера выборки. Основной минус этих методов заключается в том, что статистики, используемые в критериях, имеют асимптотическое распределение и требуют большого обьема выборки.

Существуют также байесовские оценки обьема выборки: критерий средней апостериорной дисперсии, критерий среднего покрытия, критерий средней длины и метод максимизации полезности. Первые три метода вводят функцию от обьема выборки, увеличение значений которой интерпретируется как уменьшение эффективности модели. Обьем выборки выбирается таким, при котором исследуемая функция не превышает некоторого фиксированного значения. Метод максимизации полезности максимизирует ожидание некоторой функции полезности по обьему выборки. Все эти методы опираются на апостериорное распределение, что требует достаточно большого обьема выборки.

Предлагается исследовать зависимость среднего значения логарифма правдоподобия от размера доступной выборки, а также его дисперсию при помощи метода бутстреп. После чего аппроксимировать данные две зависимости. Для вычислительного эксперимента предлагается использовать классические выборки из UCI репозитория и синтетические данные.

2 Постановка задачи

Дана выборка размера m:

$$\mathfrak{D}_m = \{x_i, y_i\}_{i=1}^m,$$

где $x_i \in \mathbb{R}^n$ - вектор признаков, $y_i \in \mathbb{Y}$ - отклик.

Предполагается, что выборка $\mathfrak{D}_m$ не противоречит гипотезе порождения данных.

Рассмотрим параметрическое семейство функций для аппроксимации неизвестного распределения $p(y|x, w)$:

$$\mathfrak{F} = \left\{ f(y, x, w) | w \in \mathbb{W}, \int_{y \in \mathbb{Y}, x \in \mathbb{R}^n} f(y, x, w) dy dx = 1 \right\}.$$

Для модели f с вектором параметров w определим функцию правдоподобия:

$$L(\mathfrak{D}_m, w) = \prod f(y, x, w).$$

Для оценки оптимального размера выборки m^* используется .... (надо определиться с критерием):

Литература

[1] *S. G. Self and R. H. Mauritsen* Power/sample size calculations for generalized linear models // Biometrics, 1988.

[2] *G. Shieh* On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000.

[3] *G. Shieh* On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005.

[4] *D. B. Rubin and H. S. Stern* Sample size determination using posterior predictive distributions // Sankhya : The Indian Journal of Statistics Special Issue on Bayesian Analysis, 1998.

[5] *Maher Qumsiyeh* Using the bootstrap for estimation the sample size in statistical experiments // Journal of modern applied statistical methods, 2002.