Секция «Теория вероятностей и математическая статистика»

Применение семплирования по Гиббсу к задаче о восстановлении многомерного распределения

Научный руководитель – Яровая Елена Борисовна

Меркушина Анна Владимировна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: anya-merkushina@yandex.ru

Восстановление пропущенных значений является актуальной проблемой, возникающей при обработке и анализе данных. Например, выборки, собранные в больших эпидемиологических исследованиях, зачастую содержат пропущенные значения [1]. При предварительной обработке данных исследователь вынужден либо восстановить пропуск в наблюдениях, либо удалить участника, имеющего потерянное значение. Замена средним, медианой или константой может оказаться некорректной (см. [2]). Также удаление всех наблюдений, связанных с утерянным, приводит к значительному сокращению объема выборки, что может негативно сказаться на точности результатов исследования. По этим причинам в последнее десятилетие были предложены усовершенствованные подходы, задача которых — восстановить совместное распределение переменных выборки и подобрать для пропущенного значения наиболее правдоподобное заполнение [2].

Целью работы является изучение восстановления пропущенных значений в реальных данных методом гиббсовского семплирования [3], а также проведение сравнительного анализа, демонстрирующего эффективность метода. Идея данного метода заключается в том, что для восстановления совместного распределения рассматриваются только условные распределения для каждой переменной. Алгоритм на каждом шаге берет случайную величину и задает ее значение при фиксированных остальных. Таким образом, последовательно моделируются n случайных величин из n одномерных условных плотностей вместо того, чтобы генерировать один n-мерный вектор за один подход с использованием полного совместного распределения. В работе показано, что последовательность генерируемых значений образует обратимую цепь Маркова, эргодическое распределение которой является искомым. Задача восстановления данных имеет ту же структуру: совместное распределение посчитать сложно, но условные распределения, как правило, известны. Таким образом, изученный алгоритм может быть использован для решения задачи восстановления данных. В работе предложен пример применения метода к задаче моделирования выборки из многомерной случайной величины, которую нельзя решить с помощью функциональных преобразований равномерной случайной величины. Для исследования использованы данные, полученные в отделе эпидемиологии ФГБУ «Национальный медицинский исследовательский центр профилактической медицины» МЗ РФ.

Источники и литература

1) Муромцева, Г. А., et al. Распространенность факторов риска неинфекционных заболеваний в российской популяции в 2012-2013гг. Результаты исследования ЭССЕ-РФ. Кардио- васкулярная терапия и профилактика 13(6) (2014).
2) Buuren, S. V., Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. Journal of statistical software (2010).
3) Murphy, K. P. Machine learning: a probabilistic perspective. MIT press (2012).