Секция «Теория вероятностей и математическая статистика»

Ослабление предположения линейности при анализе главных компонент. Метод t - SNE и его применения.

Ивлев Олег Евгеньевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: olivlegerr@gmail.com

В машинном обучении при анализе данных в пространствах высокой размерности может возникать ряд сложностей, которые называют “проклятием размерности” [1]. В таких ситуациях принято использовать методы снижения размерности, то есть преобразование данных, состоящее в уменьшении числа переменных и переходе к новым. Предложенные методы не нашли широкого применения в популяционных исследованиях до последних лет в силу того, что они не были реализованы в статистических пакетах до настоящего времени. Мы остановимся на изучении методов снижения размерности при наличии скрытых переменных [2] и их применении в эпидемиологическом исследовании. В работе сравниваются анализ главных компонент (PCA) [3], метод главных кривых (principal curves) [4] и стохастическое вложение соседей с t-распределением (t- SNE) [5], предложенные в 1901, 1989 и 2008 гг., соответственно. Для каждого метода в докладе представлено его краткое теоретическое описание, особенности применения и практическая реализация. Цель работы - проверить гипотезу о существовании двумерного многообразия, вложенного в пространство биомаркеров высокой размерности. Причем предполагается, что внутренняя структура многообразия определяется переменной, которая не используется в алгоритмах снижения размерности. Для проверки гипотезы с помощью метода главных кривых строится статистика, основанная на сравнении результатов линейного и нелинейного методов снижения размерности. На модельных примерах удалось выделить внутреннюю структуру двумерного многообразия, вложенного в трехмерное пространство. Однако для реального набора биомаркеров таким способом выделить многообразие не удалось. При этом привлечение обобщенных аддитивных моделей позволило выявить наличие нелинейной связи между переменной (возраст), неиспользуемой в алгоритмах снижения размерности, и изучаемыми биомаркерами. Таким образом, мы предполагаем, что связь между биомаркерами обладает слишком малой нелинейностью, чтобы она могла быть выявлена с помощью методов снижения размерности. Все перечисленные методы были применены к реальным данным, полученным в отделе эпидемиологии хронических неинфекционных заболеваний ФГБУ «Государственный научно-исследовательский центр профилактической медицины» МЗ РФ.

Источники и литература

1) Bellman R. Dynamic programming, 1957 //A very comprehensive reference with many economic examples is. – 2003.
2) Theodoridis S. Machine learning: a Bayesian and optimization perspective. – Academic press, 2015.
3) Pearson K. LIII. On lines and planes of closest fit to systems of points in space //The London, Edinburgh, and Dublin philosophical magazine and journal of science. – 1901. – Т. 2. – №. 11. – С. 559-572.

4) Hastie T., Stuetzle W. Principal curves //Journal of the American Statistical Association. – 1989. – Т. 84. – №. 406. – С. 502-516.

5) Van der Maaten L., Hinton G. Visualizing data using t-SNE //Journal of machine learning research. – 2008. – Т. 9. – №. 11.