Секция «Теория вероятностей и математическая статистика»

Некоторые алгоритмы регрессионного анализа и особенности их применения.

Научный руководитель – Яровая Елена Борисовна

Савинов Эдуард Эдуардович
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: eddy.savinovcm@gmail.com

В регрессионном анализе одной из основных проблем является отбор признаков, вносящих статистически значимый вклад в изучаемую модель. Существует множество алгоритмов, позволяющих осуществлять такой отбор. В работе мы остановимся на применении следующих алгоритмов: прямой и обратной пошаговой регрессии (stepwise regression,[1]), регрессионном методе наименьших углов (LARS, least angle regression,[2]) и лассо регрессии (LASSO, Least absolute shrinkage and selection operator,[3]). Цель работы - сравнить особенности применения перечисленных методов отбора признаков и продемонстрировать их работу на реальных данных. Алгоритм прямой пошаговой регрессии основан на последовательном включении переменных в регрессионное уравнение. Порядок включения определяется коэффициентом корреляции невключенных переменных и целевой переменной, с поправкой на уже включенные переменные, а так же проверкой гипотез о значимости переменных. В случае обратной пошаговой регрессии из модели, содержащей все переменные, происходит исключение переменных посредством проверки гипотез об их значимости. В отличие от пошаговой регрессии, алгоритм LARS, вместо последовательного добавления независимых переменных, на каждом шаге изменяет их веса. Веса изменяются так, чтобы можно было получить наибольшую корреляцию с вектором остатков регрессии. Основным достоинством LARS по сравнению с пошаговой регрессией является то, что он выполняется за число шагов, не превышающих числа независимых признаков в модели. Частным случаем алгоритма LARS является алгоритм LASSO. Алгоритм LASSO основан на идее введения ограничений на абсолютную величину коэффициентов регрессии, данная модификация позволяет бороться с мультиколлинеарностью. Модели применяются к данным, полученным в Российском кардиологическом научно производственном комплексе Минздрава РФ. Качество моделей определяется на основе коэффициента детерминации и среднеквадратичной ошибки.

Источники и литература

1) Hocking, R. R. (1976) "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32.
2) Efron B., Hastie T., Johnstone J., Tibshirani R. Least Angle Regression. //The Annals of Statistics.2004 Vol.32, № 2,p.407-499.
3) Tibshirani R. Regression shrinkage and Selection via the Lasso. //Journal of the Royal Statistical Society.Series B(Metodological). 1996 Vol. 32, № 1, p.267-288.