

O rzadkiej predykcji na podstawie danych wysokiego wymiaru

Piotr Pokarowski

Uniwersytet Warszawski

Streszczenie

Dane wysokiego wymiaru, w których liczba predyktorów jest dużo większa od liczby obserwacji są powszechne w nauce, biznesie czy w służbach publicznych. Często celem analizy takich danych jest znalezienie rzadkiego/oszczędnego (sparse/parsimonious) modelu predykcyjnego, czyli prostej i jednocześnie silnej zależności między małą grupą predyktorów, a zmienną zależną y . Główną trudnością takich zadań jest wzajemne skorelowanie predyktorów, co powoduje, że metody przesiewu wykorzystujące tylko informację o związku pojedynczego predyktora z y nie rozwiązują problemu. Znaczącą poprawę można uzyskać dzięki metodom opartym na maksymalizacji wiarygodności z karą indukującą oszczędność, takim jak Lasso.

Wykład będzie poświęcony algorytmowi Lasso, jego modyfikacjom oraz algorytmom zmniejszającym zbiór predyktorów otrzymanych przez Lasso bez istotnego pogorszenia jakości predykcji.