Projekt 1 – Algorytmy Statystyki Praktycznej

8 marca 2019

Rozważmy model liniowy

$$Y = X\beta + \varepsilon \,,$$

gdzie $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ oraz $\varepsilon \in \mathbb{R}^n$. Jest to model z wyrazem wolnym, czyli pierwsza kolumna macierzy X zawiera same jedynki. W sytuacji gdy $p > n$ estymacja β wymaga regularyzacji, dodania kary. Jedną z najpopularnijeszych metod jest LASSO, czyli estymator $\hat{\beta}$ jest rozwiązaniem problemu optymalizacyjnego

$$\min_{\beta}\{\|Y - X\beta\|^2 + \lambda \sum_{i=2}^{p} |\beta_i|\}$$

Uwaga: β_1 odpowiadająca wyrazowi wolnemu nie jest karana

Zadanie 1
Jedną z możliwości rozwiązania powyższego problemu optymalizacyjnego jest algorytm ADMM. Jego opis można znaleźć na stronie `http://web.stanford.edu/~boyd/admm.html` oraz `https://web.stanford.edu/~boyd/papers/pdf/admm_distr_stats.pdf`. Pierwsza część projektu polega na napisaniu funkcji która rozwiązuje problem LASSO za pomocą ADMM. Funkcja powinna przyjmować wektor Y, macierz X oraz ilość λ (Ponieważ nie ma możliwości wybrania dobrego parametru λ rozwiązuje się problem dla siatki kar a ostateczną wybiera sie za pomocą walidacji krzyżowej) oraz parametr ρ algorytmu ADMM. Funkcja jako wynik powinna zwracać macierz wyestymowanych β oraz wektor λ.

Funkcja powinna przeskalować macierz X tak aby sumy kwadratów kolumn były równe 1. Wektor λ powinien być otrzymany w następujący sposób λ_1 powinna być najmniejsza możliwa jaka daje pusty model (tylko $\beta_1 \neq 0$). (Można ją łatwo policzyć analitycznie). Kolejne λ_i powinny tworzyć ciąg geometryczny łączący λ_1 z $\lambda_k = 0.001\lambda_1$, gdzie k jest długością.

Zadanie 2
Polega na porównaniu swojej implementacji z implementacjami dostępnymi w python. W tym celu należy wygenerować dane i przetestować algorytmy. Należy wygenerować dane dla $n = 100$ i $p = 10000$, gdzie macierz X generowaną z $N(0.\Sigma)$, gdzie

1. $\Sigma = Id$ (niezależne)

2. Σ macierzą autokorelacji czyli $\Sigma_{ij} = \varrho^{|i-j|}$ dla $\varrho = 0.6, 0.9$

3. Σ parami tak samo skorelowane $\Sigma_{ij} = \varrho$ dla $i] \neq j$ oraz $\Sigma_{ii} = 1$ ponownie $\varrho = 0.6, 0.9$.

Wektor β powinien być równy 2 dla $k = 5, 10, 100$ losowo wybranych współrzędnych, pozostałe zerowe. Szum ε powinien być wygenerowany ze standardowego rozkładu normalnego.

W symulacje powinny określić odpowiednią wartość parametru ρ algorytmu ADMM. Oraz porównać szybkość/dokładność własnej implementacji ze standardowymi implementacjami.