

- Załóżmy, że mamy n par $(x_i, y_i), i = 1, \dots, n$, gdzie $x_i = (x_{i1}, \dots, x_{id}) \in \mathbf{R}^d$ (cechy - wektor kolumnowy) oraz $y_i \in \{0, 1\}$ (klasy). Załóżmy, że y_i są realizacjami zmiennej losowej Y_i o rozkładzie 0-1, który to rozkład zależy od x_i . Oznaczmy

$$p_i = P(Y_i = 1 | x_i).$$

- Cel: znaleźć “najlepsze” oszacowania na $p_i, i = 1, \dots, n$.
- Załóżmy (model regresji logistycznej), że

$$\log \left(\frac{p_i}{1 - p_i} \right) = x_i \beta^T,$$

gdzie $\beta = (\beta_1, \dots, \beta_n)$.

- Chcemy de facto znaleźć β , bo wtedy $p_i = \frac{\exp(x_i \beta^T)}{1 + \exp(x_i \beta^T)}$.
- Funkcja wiarygodności

$$\mathcal{L}(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} = \prod_{i=1}^n \left(\frac{\exp(x_i \beta^T)}{1 + \exp(x_i \beta^T)} \right)^{y_i} \left(1 - \frac{\exp(x_i \beta^T)}{1 + \exp(x_i \beta^T)} \right)^{1-y_i}$$

- Można wyliczyć, że

$$f(\beta) := -\ln(\mathcal{L}(\beta)) = \sum_{i=1}^n \{ \log(1 + \exp(x_i \beta^T)) - y_i x_i \beta^T \} =: \sum_{i=1}^n f_i(\beta).$$

Zauważmy, że funkcje $f_i(\beta), i = 1, \dots, n$ są wypukłe.

W sytuacji, gdy n jest duże koszt jednego kroku algorytmu optymalizacji jest drogi (zależy liniowo od n). W celu zmniejszenia kosztu stworzony został szereg metod opartych na losowaniu jednego składnika sumy w każdym z kroków.

W pracy <https://arxiv.org/abs/1407.0202> zaproponowano jedną z takich metod SAGA oraz porównano ją z konkurencyjnymi metodami między innymi SVRG, Finito. Celem projektu jest zbadania czy te metody poprawiają czas działania algorytmu w porównaniu do standardowych metod.

Zadania:

1. Ustal d i w miarę duże n , wylosuj macierz X wymiaru $n \times d$ z wierszami $x_i, i = 1, \dots, n$, ($x_i = (x_{i1}, \dots, x_{id})$) o rozkładzie normalnym o średniej 0 i macierzy kowariancji Σ . Najczęściej dla danych symulowanych używa się trzech typów macierzy Σ : $\Sigma = Id$, parami tak samo skorelowane $\sigma_{ii} = 1, \sigma_{ij} = \rho < 1$, autokorelacja $\sigma_{i,j} = \rho^{|i-j|}$. Następnie wylicz $p_i = \frac{\exp(x_i \beta^T)}{1 + \exp(x_i \beta^T)}, i = 1, \dots, n$ oraz wysymuluj $Y_i \sim \text{Bin}(1, p_i)$. Zastosuj wspomniane metody do estymacji p_i . Porównaj wyestymowane wartości z rzeczywistymi.
2. Powtórz eksperymenty dla modelu gdzie większość $\beta_i = 0$ (np. każda współrzędna β_i jest albo zerem z prawd. 0.8, albo z pozostałym prawd. jest losowa z jakiegoś zbioru spoza $(-1, 1)$) korzystając z modelu regresji logistycznej z karą LASSO czyli, szukamy β minimalizującej

$$-\ln(\mathcal{L}(\beta)) + \lambda \sum_{i=1}^d |\beta_i|.$$

Dobrze by było odpowiedzieć jak wybór metody zależy od n oraz jak zależy od pożądanej dokładności rozwiązania (typową regułą stopu jest zatrzymanie algorytmu jak zmiana w określonej liczbie kroków jest dostatecznie mała).

- Projekt można wykonać **w grupach do 3 osób**.
- Należy oddać **kod** napisany w języku **python** oraz **raport** w formacie **.pdf** opisujący uzyskane wyniki.
- **Termin oddania: 25.03.2018**