Theoretical task 2

due January 21.

Remark: all solutions should be short, mathematically precise and contain proof unless qualitative explanation / intuition is needed. Solutions should be sent electronically to *ml.tasks@yandex.ru* and can be written in any clear and understandable format - latex, handwritten/scanned or other. Late submissions (by no more than 3 days) will be penalized by 50%, identical solutions will not be graded. The title of your e-mail should be "ICL homework <homework number> - <your first name and last name> ".

1. Derive analytical solution for weighted regression:

$$\sum_{n=1}^{N} w_n \left(x_n^T \beta - y_n\right)^2 \to \min_{\beta \in \mathbb{R}}$$

in terms of matrix of weights diagonal matrix $W = diag\{w_1, ...w_N\} \in \mathbb{R}^{NxN}$, desigm matrix $X \in \mathbb{R}^{NxD}$ and outputs vector $Y \in \mathbb{R}^{Nx1}$, where D is the number of features.

2. Solution for ridge regression is

$$\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

where $X \in \mathbb{R}^{NxD}$ is the design matrix, $Y \in \mathbb{R}^{Nx1}$ is vector of outputs,$\lambda > 0$ is user specified regularization multiplier. Prove that $X^T X + \lambda I \in \mathbb{R}^{DxD}$ is always full rank (and thus invertible) by showing that for any non-zero vector $v \in \mathbb{R}^{Dx1}$

$$\left(X^T X + \lambda I\right) v \neq \mathbf{0} \quad \text{(never equals to zero vector)}$$

3. Write stochastic gradient descent for minimatch size $K = 1$ and gradient descent algorithm for the following cases:

   (a) classification with Perceptron loss $\mathcal{L}(M) = [-M]_+$ using indicator notation $\mathbb{I}[condition] = \begin{cases} 1, & \text{if condition is satisfied} \\ 0, & \text{otherwise.} \end{cases}$

   (b) classification with exponential loss$\mathcal{L}(M) = e^{-M}$.