Theoretical task 2

due 9:00 February 1 (Monday).

To simplify notation $g(x) = w^T x + w_0$ for linear methods suppose that x is augmented with new unity feature $x \leftarrow [1;\, x]$ and accordingly $w \leftarrow [w_0; w]$. So now we can write $g(x) = \langle w, x \rangle$.

1. For binary classification $y \in \{+1, -1\}$ we may define a linear classifier with the following pair of discriminant functions: $g_{+1}(x) = \langle w_{+1}, x \rangle$ and $g_{-1}(x) = \langle w_{-1}, x \rangle$, or equivalently with the following decision rule $\widehat{y}(x) = sign[\langle w, x \rangle]$ for $w = w_{+1} - w_{-1}$. What is the connection between two definitions of margin presented on lectures [here y is the correct class for x]:

   (a) $M(x, y) = g_y(x) - \max_{c \neq y} g_c(x)$

   (b) $M(x, y) = y \langle w, x \rangle$

2. For Perceptron of Rosenblatt method $\mathbb{I}[M < 0] \approx \mathcal{L}(M) = [-M]_+ = max\{-M, 0\}$ and for logistic regression $\mathbb{I}[M < 0] \approx \mathcal{L}(M) = \ln(1 + e^{-M})$ [see page 30 of lecture on linear classifiers for details]. For both methods:

   (a) Plot $\mathcal{L}(M)$ on the same graph

   (b) Plot $\frac{\partial \mathcal{L}(M)}{\partial M}$ on the same graph

   (c) Write down the update rule of weights for stochastic gradient descent method

   (d) Looking at the results of a), b) and c), what is the qualitative difference between the two methods?

   (e) Write down the update rule of weights for gradient descent method.

   (f) What is the advantage of c) compared to e) update formula?