Theoretical task 2

due January 31 9:00 (Tuesday).

Remark: solutions should be **given in printed or written form** on the lecture on January 31. All solutions should be short, mathematically precise and contain proof unless qualitative explanation / intuition is needed. Late solutions should be sent to *v.v.kitov@yandex.ru* and can be written in any clear and understandable format - latex, handwritten/scanned or other. Late submissions 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. Suppose your training set consists of N samples and you generate bootstrap pseudosample of the same size.
   (a) What is the probability, that a particular observation (object) will not appear in the whole bootstrap pseudosample?
   (b) What is the limit of this probability as $N \to \infty$?

2. Suppose we project feature vectors $x_1, x_2, ...x_N$ on linear subspace of lower dimension $K \leq D$, so that the projection of x_n is p_n and orthogonal complemement (or equivalently speaking error of approximation) is $h_n = x_n - p_n$, $n = 1, 2, ...N$. Suppose, that by looking at all possible subspaces of given dimensionality K we select the subspace so that the squared sum of L_2 norms of orthogonal complements is minimized:

$$\|h_1\|^2 + \|h_2\|^2 + \|h_N\|^2 \to \min$$

   Prove that this is equivalent to maximizing the squared sum of L_2 projections:

$$\|p_1\|^2 + \|p_2\|^2 + \|p_N\|^2 \to \max$$

3. Consider M classifiers $f_1(x), ...f_M(x)$, performing binary classification. Suppose each of the models makes mistakes independently with probability $p < 0.5$. Prove that probability of incorrect classification by majority voting $p(\text{incorrect } y|x) \xrightarrow{M\to\infty} 0$.

   Hint: you may make use of central limit theorem.

4. Suppose, you perform binary classification with score of the positive class, compared to the score of the negative class being equal to the discriminant function $g(x) = w^T x$ and classification made by the rule $\widehat{y}(x) = sign(w^T x)$. Suppose that to measure positive class probability you use heuristics $p(y = +1|x) = \sigma(w^T x)$, where $\sigma(u) = 1/(1 + e^{-u})$ is so called sigmoid function.
   (a) explain why maximum posterior probability classifier $\widehat{y}(x) = \arg\max_{y\in\{+1,-1\}} p(y|x)$ will give the same classes as $\widehat{y}(x) = sign(w^T x)$
   (b) estimation of w using maximum likelihood estimation $\widehat{w} = \arg\max_w p(y_1, ...y_N|x_1...x_N)$, given that all objects are independently and identically distributed, is equivalent to finding w with logistic loss minimization:

$$\widehat{w} = \arg\min_w \sum_{n=1}^{N} \mathcal{L}(M_n), \quad \mathcal{L}(M) = \ln(1 + e^{-M}),\ M_n = w^T x_n y_n$$

   Hint: you may use sketch of proof of (b) in the "logistic regression" section of lecture slides about linear classifiers. You need to write down all the details of the proof.