Theoretical task 1

due January 28 23:59 (Saturday).

Remark: all solutions should be short, mathematically precise and contain proof unless qualitative explanation / intuition is needed. 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. Consider real numbers $z_1, z_2, ...z_N$. Find such constant approximation μ of these numbers, so that
   (a) the sum of square deviations from these points to μ $\sum_{n=1}^{N}(z-\mu)^2$ is minimized.
   (b) the sum of absolute deviations from these points to μ $\sum_{n=1}^{N}|z_n-\mu|$ is minimized.

   Hint: will the functions by convex? why? you may look at the derivative of the minimized criterion.
2. Suppose, we have some fixed classifier, specified by some function $f: x \to y$. Are discriminant functions $g_1(x), ...g_C(x)$ for this fixed classifier defined uniquely or we can select multiple sets of discriminant functions, which yield the same predictions for all x?
3. Consider ridge regression:
$$\sum_{n=1}^{N}\left(x_n^T\beta - y_n\right)^2 + \lambda\sum_{d=1}^{D}\beta_d^2 \to \min_{\beta}$$
   (a) Derive the formula for optimal β.
   (b) Explain qualitatively, why the problem of ambiguity does not arise when features are correlated, but we impose L_2 regularization in target criterion?
4. Under what selection of function $K(u)$ and window width $h(x)$ will Parzen window classifier turn exactly into K-nearest neighbors method?
5. Suppose that you have a binary random classifier, assigning probabilities
$$\begin{aligned} p(y=+1|x) &= \xi \\ p(y=-1|x) &= 1-\xi \end{aligned}$$
   where ξ is a random variable uniformly distributed on $[0,1]$ independent of x.
   (a) Suppose you assign class when $p(y=+1|x) \geq \mu$ for some threshold μ. What will be $TPR(\mu)$ and $FPR(\mu)$?
   (b) Plot the ROC curve for this classifier.
6. Write down update rules for weights in case of linear classifier optimization with logistic loss $\mathcal{L}(M) = \mathcal{L}(x,y,w) = \ln\left(1+e^{-w^T xy}\right)$ in
   (a) gradient descent method
   (b) stochastic gradient descent method

   Your solutions should depend only on (perhaps functionally transformed) feature vectors, correct answers and previous estimate of w.