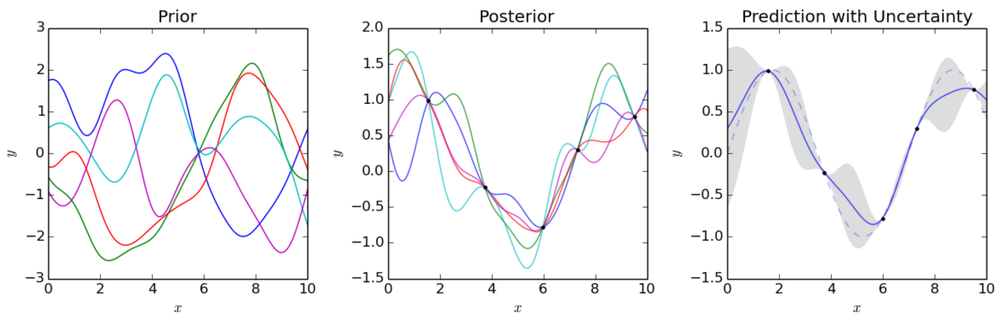
Gaussian Processes for Active Learning in Classification

Active Learning – is the new way to learn models based on adaptive increasing training data set with previously unlabeled points. Usually point choice uses estimation of uncertainty of the model at this point. Recently new criterion for active learning was proposed in the work Active Learning in the Overparameterized and Interpolating Regime by Mina Karzand and Robert D. Nowak. The criterion estimates norm of the function that describes training set + one unlabeled point. Point that caused the maximum norm of the function is considered to be “the hardest” for the model to describe and the most useful to learn. **Purpose** of the work is to compare different norms in the criterion and see how good they are – there was no similar work before.

The criterion was developed for overparameterized neural networks, however learn such network is quite costly and get analytical expression for the criterion is not that simple. That’s why in this work we use Gaussian processes – stochastic processes (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution. The intuition under similarity of Gaussian process and overparametrized neural network is: the output of the one layer of neural network is linear combination of inputs. If we suppose that inputs (or features) are independent and the number of weights in the layer is big enough, then works central limit theorem, that states: sum of the big number of independent random variables has distribution close to the normal one. And any finite combination of Gaussian process’s variable has normal distribution.



An example of how Gaussian process regression works

Recently (May 2019) in the work Active Learning in the Overparameterized and Interpolating Regime by Mina Karzand and Robert D. Nowak the new criterion for active learning was proposed.
Let $\mathcal{L} = (x_1, ..., x_n)$ – training data with labels (y_1, y_n) and $\mathcal{U}$ – set of examples models doesn’t know labels for. Let f – function, interpolating training dataset, with minimum norm (this condition is gotten from the experiments – such models usually have good interpolating properties. Let’s define $f_t^u(x)$ - as minimum-norn function that interpolates training data combined with the point $u \in \mathcal{U}$ with label t. The label $t(u)$ we will choose by one of these ways:

$$t^{(1)}(u) = argmin_{t\in\{-1,1\}}\|f_t^u(x)\|,\ t^{(2)}(u) = \begin{cases} +1 \text{ if } f(u) \geq 0, \\ -1 \text{ if } f(u) < 0 \end{cases} \quad (1)$$

Defined $t(u)$, let $f^u(x) = f^u_{t(u)}(x)$. Introduce also *score*-functions:

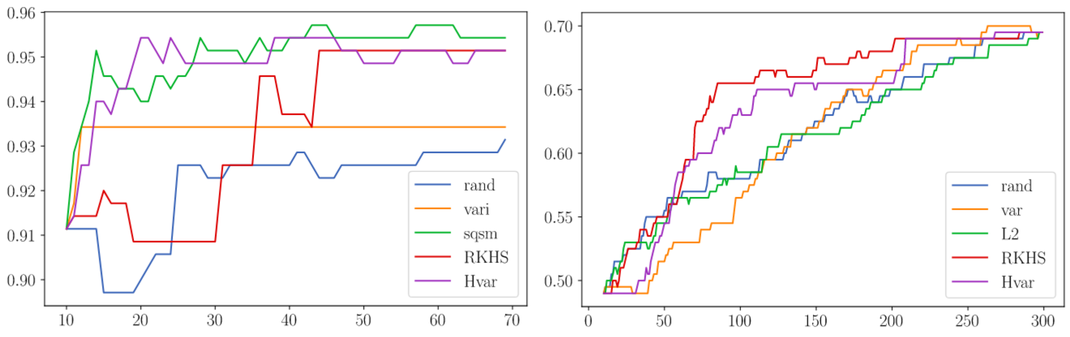
$$score^{(1)}(u) = \|f^u(x)\|,\ score^{(2)}(u) = \|f^u(x) - f(x)\|, \quad (2)$$

In the first case the most *score* get the least smooth function, in the second case – the function, that differs the most from previous one. We expect that point with the largest *score* is the most imformative.
Then the next point for labeling is

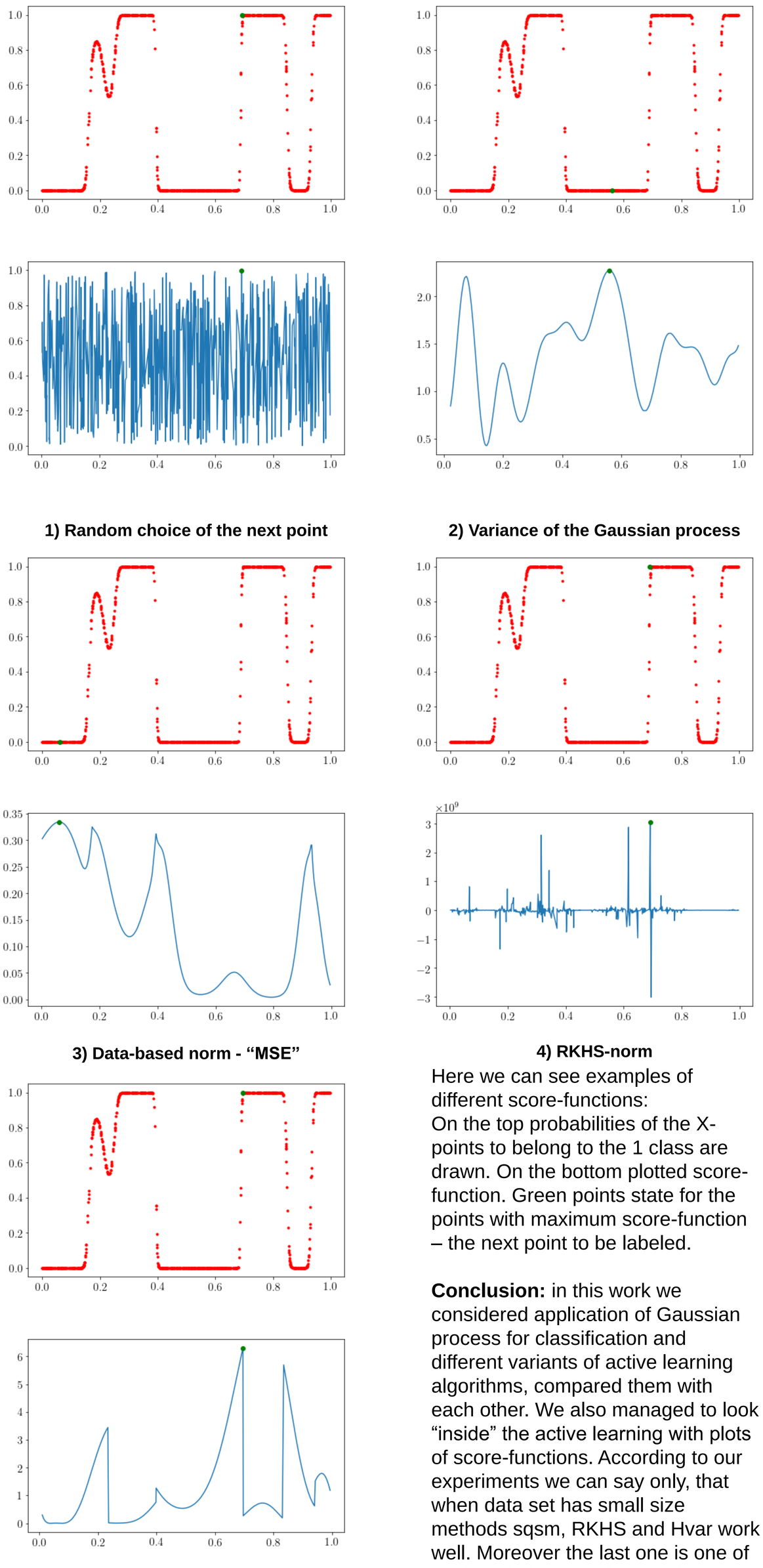
$$u^* = argmax_{u\in\mathcal{U}} score(u).$$

In (1) и (2) we intentionally do not define specific norm for variety of *score*-functions. If these norms are the same and we chose the first *score*-function, then the new point u^* can be determined by

$$\|f^{u^*}(x)\| = max_{u\in\mathcal{U}} min_{t\in\{-1,1\}}\|f_t^u(x)\|. \quad (3)$$



There are two plots for accuracy of the classification on the test data for different methods of getting the new point. On the left the plot is for synthetic data with dimensionality = 1. On the right – also synthetic data, but with dimensionality = 5. The discreteness of the accuracy is caused by small test size – unfortunately, Gaussian processes are quite slow due to matrix inversion, because of limited time we had to use small data set.



1) Random choice of the next point

2) Variance of the Gaussian process

3) Data-based norm - “MSE”

4) RKHS-norm

5) RKHS-norm * variance

Here we can see examples of different score-functions:
On the top probabilities of the X-points to belong to the 1 class are drawn. On the bottom plotted score-function. Green points state for the points with maximum score-function – the next point to be labeled.

Conclusion: in this work we considered application of Gaussian process for classification and different variants of active learning algorithms, compared them with each other. We also managed to look “inside” the active learning with plots of score-functions. According to our experiments we can say only, that when data set has small size methods sqsm, RKHS and Hvar work well. Moreover the last one is one of the best in both made experiments. To get more informative results obviously we need more experiments with much bigger sets of data (and real data also!).

M9

Kotova Daria (MIPT), Maxim Panov (Skoltech)

ITAS 2019