Gaussian Process in Active Learning for Classification

Daria Kotova, Maxim Panov

April 9, 2020

Abstract

There is a working file with description of what we are doing.

1 Gaussian Process

In this section we briefly review the definition of a Gaussian process and some important equations connected with it. A comprehensive overview of Gaussian processes is presented in [2].
Gaussian process $f(x)$ – stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution. Here $x \in \mathbb{R}^D$, where D - arbitrary natural number.
Gaussian process is defined by mean and covariance functions. Often mean function is chosen as constant 0. We introduce the notation:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] = 0, \\ K(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \end{aligned}$$

Regression. We want to get posterior on process $f(x)$, having observed points X. Prior is defined by:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right),$$

where $*$ corresponds to the test set. With Bayes formula and some calculations we show that posterior will look like:

$$(f_*|X, X_*, f) \sim \mathcal{N}(\hat{f}, \hat{\sigma}^2), \tag{1}$$

where $\hat{f} = K(X_*, X)K(X, X)^{-1}f$ is the posterior mean function,
$\hat{\sigma}^2 = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*))$ is the posterior covariance function.

The reader can notice that interpretation of this result for covariance function is quite intuitive: $K(X_*, X_*)$ is prior covariance and positive term, corresponding to new information, is substracted from prior knowledge.
To sum up, in regression case Gaussian process give us not only mean, but also variance – the measure of uncertainty in a giving point. This property is advantage of Gaussian proccess. However, we should point out that matrix inversion in (1) is extremely time consuming.

2 Active Learning

Now let's move on to basics of active learning approach. It is based on an assumption that model will give better results having less training points, if we allow it to choose points to train on by itself.
The algorithm is quite simple: the model chooses the next point which label it wants to get. Then ask an "oracle" for the label and somehow incorporates new knowledge. Then it repeats all the steps till some condition for stopping. Review of active learning can be found in [4].
We will consider pool-based sampling, which assumes that there is a small set of labeled data $\mathcal{L} = (x_1, ..., x_n)$ and a large pool of unlabeled data $\mathcal{U}$ available. New points are selectively drawn from the pool taking into account some acquisition function $g(u)$. Often new point u^* delivers maximum of $g(u)$. This approach is also illustrated in Fig.1.
In this work we tried different $g(u)$. As a reference we used the work [3] where general criterion

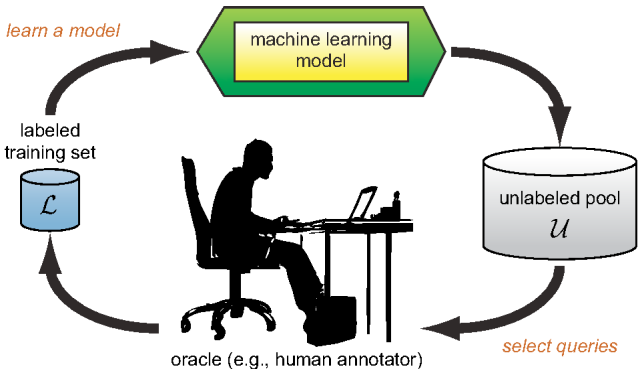


Figure 1: Illustration of active learning algorithm using pool-based method.

and some specific variants were introduced. Now we will consider results of this work.
Let f be the minimum norm function that interpolates labeles examples. Define $f_t^u(x)$ is the minimum norm interpolating function based on $\mathcal{L}$ and the point $u \in \mathcal{U}$. To get $f_t^u(x)$ we suppose that u has label t (since in supervised learning we have to have labeles for put points). Then our $g(u)$ can be presented as (both variants are possible):

$$g(u) = \|f_t^u(v)\| \text{ or } g(u) = \|f_t^u(v) - f(v)\|.$$

3 Criteria

Here we compare 6 different ways to choose new point u^* from unlabeled data $\mathcal{U}$. Having assumed that $u^* = \underset{u \in \mathcal{U}}{argmax}(g(u))$, we change $g(u)$ and compare results.

1. *Random* – just to check if results are adequate (rand).
2. *Variance* – new point corresponds to the maximum of variance:

$$g(u) = \hat{\sigma}^2(u) \text{ (mvar)}.$$

3. *2-norm* – the criterion was introduced in [3]. A new point is the argmax of:

$$g(u) = \|f_t^u(v) - f(v)\|_{\mathcal{U}} = \sqrt{\sum_{v \in \mathcal{U}} (f_t^u(v) - f(v))^2} \text{ (sqsm)}.$$

4. *RKHS-norm* – this criterion was also introduced in [3]. However, we want to provide more comprehensive inference here:

$$g(u) = \|f_t^u(v)\|_{\mathcal{H}} = \vec{\widetilde{y}_t^T} \widetilde{K} \vec{\widetilde{y}_t} \text{ computed with RKHS-norm (RKHS)}.$$

Let's infer the $\|f_t^u(v)\|_{\mathcal{H}}$. Let $\widetilde{K} = \begin{pmatrix} K & \vec{a} \\ \vec{a}^T & b \end{pmatrix}$, where $b = K(u,u), \vec{a} = [K(x_1,u), ..., K(x_n,u)]^T$. Using Schur complement formula we have $\widetilde{K}^{-1}$ as:

$$\begin{aligned} \widetilde{K}^{-1} &= \begin{pmatrix} K^{-1} + K^{-1}\vec{a}(b - \vec{a}^T K^{-1}\vec{a})^{-1}\vec{a}^T K^{-1} & -K^{-1}\vec{a}(b - \vec{a}^T K^{-1}\vec{a})^{-1} \\ -(b - \vec{a}^T K^{-1}\vec{a})^{-1}\vec{a}^T K^{-1} & (b - \vec{a}^T K^{-1}\vec{a})^{-1} \end{pmatrix} \\ &= \begin{pmatrix} K^{-1} + \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}} & \frac{-K^{-1}\vec{a}}{b - \vec{a}^T K^{-1}\vec{a}} \\ \frac{-\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}} & \frac{1}{b - \vec{a}^T K^{-1}\vec{a}} \end{pmatrix}. \end{aligned}$$

Then $\|f_t^u(v)\|$ turns into:

$$\begin{aligned} \|f_t^u(v)\| &= \vec{y}^T K^{-1}\vec{y} + \vec{y}^T \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}}\vec{y} - 2t\vec{y}^T K^{-1}\frac{\vec{a}}{b} - 2t\vec{y}^T \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}}\frac{\vec{a}}{b} + \frac{1}{b - \vec{a}^T K^{-1}\vec{a}} \\ &= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a}\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} - 2t\frac{\vec{y}^T K^{-1}\vec{a}}{b}\left(1 + \frac{\vec{a}^T K^{-1}\vec{a}}{b - \vec{a}^T K^{-1}\vec{a}}\right) + \frac{t^2}{b - \vec{a}^T K^{-1}\vec{a}} \\ &= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a}\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} - \frac{2t\left(\vec{y}K^{-1}\vec{a}\right)}{b - \vec{a}^T K^{-1}\vec{a}} + \frac{t^2}{b - \vec{a}^T K^{-1}\vec{a}} \\ &= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a} - t\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} = \vec{y}^T K^{-1}\vec{y} + \frac{(f(u) - y(u))^2}{b - \vec{a}^T K^{-1}\vec{a}}. \end{aligned}$$

That means

$$g(u) = \|f_t^u(v)\|_{\mathcal{H}} = \vec{\widetilde{y}_t^T} \widetilde{K} \vec{\widetilde{y}_t} = \|f(v)\|_{\mathcal{H}} + \frac{(1 - t \cdot f(u))^2}{b - \vec{a}^T K^{-1}\vec{a}},$$

It is important to emphasize that denominator here is the posterior variance of the model at the point u (see (1)).

5. *RKHS-norm · variance* – RKHS-norm turns out to have too huge values. Multiplying by variance allows solve the problem and also make quality of the criterion better. This idea was introduced in [1] (Hvar).

$$g(u) = (\|f_t^u(v)\| - \|f(v)\|_{\mathcal{H}}) \cdot \hat{\sigma}^2(u) = (1 - t \cdot f(u))^2.$$

6. *2-norm in formula* – in sqsm we directly learned new model for each u. However, it can be simplified using previous knowledge (l2fm):
   Let $\widetilde{a} = \begin{pmatrix} \vec{a} \\ K(u,v) \end{pmatrix}$. Where u is the new point and $v \in \mathcal{U}$ is the point we calculate $f(v)$ for.
   Let's compute difference $f_t^u(v) - f(v)$:

$$\begin{aligned} f_t^u(v) - f(v) &= \vec{\widetilde{a}}^T \widetilde{K}^{-1} \vec{\widetilde{y}} - \vec{a}^T K^{-1} \vec{y} \\ &= \begin{pmatrix} \vec{a}^T & K(u,v) \end{pmatrix} \begin{pmatrix} K^{-1} + \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}} & \frac{-K^{-1}\vec{a}}{b - \vec{a}^T K^{-1}\vec{a}} \\ \frac{-\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}} & \frac{1}{b - \vec{a}^T K^{-1}\vec{a}} \end{pmatrix} \begin{pmatrix} \vec{y} \\ t \end{pmatrix} - \vec{a}^T K^{-1} \vec{y} \\ &= \frac{(\vec{a}^T K^{-1}\vec{a} - K(u,v))(\vec{a}^T K^{-1}\vec{y} - t)}{b - \vec{a}^T K^{-1}\vec{a}}. \end{aligned}$$

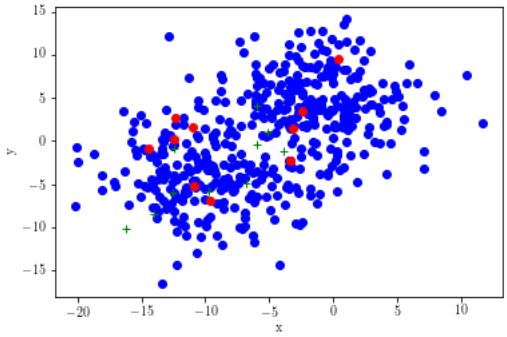
   Finally, we get:

$$\begin{aligned} g(u) &= \|f_t^u(v) - f(v)\|_{\mathcal{U}} = \sqrt{\sum_{v \in \mathcal{U}} (f_t^u(v) - f(v))^2} \\ &= \sqrt{\sum_{v \in \mathcal{U}} \left(\frac{(\vec{a}^T K^{-1}\vec{a} - K(u,v))(\vec{a}^T K^{-1}\vec{y} - t)}{b - \vec{a}^T K^{-1}\vec{a}} \right)^2}. \end{aligned}$$

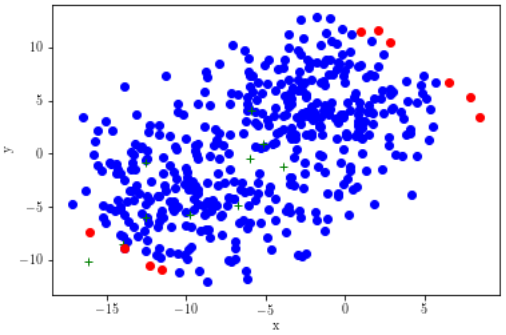
4 Experiments

4.1 2 blobs

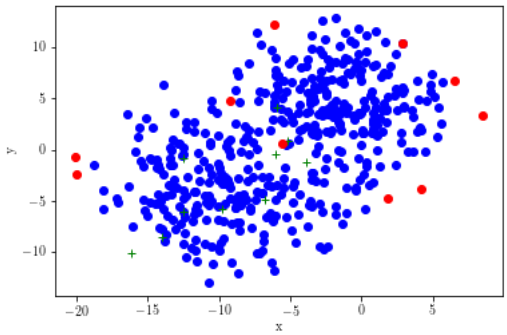
We took 2 blobs of 2-dimensional points that consist of 1000 points total. 500 of them went to the test set and training size was changing from 20 to 500 points. Figures 2, 3, 4 show what points different score-functions tend to choose for cases when in the training set are 50, 100 and 150 points.



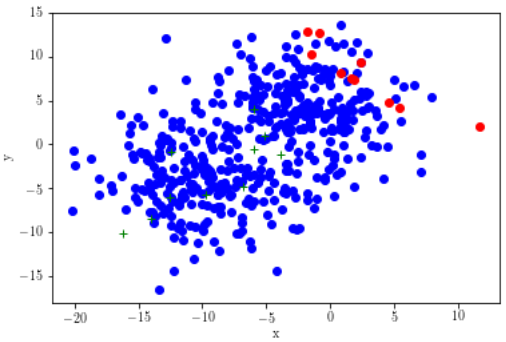
Random



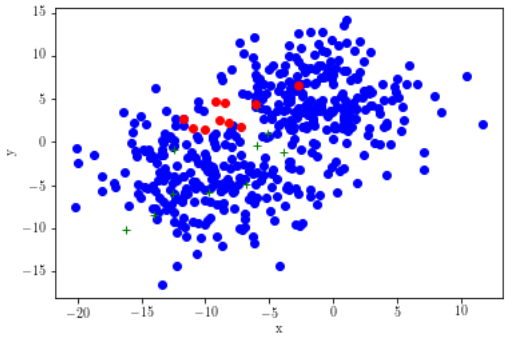
Max-variance



2-norm in formula

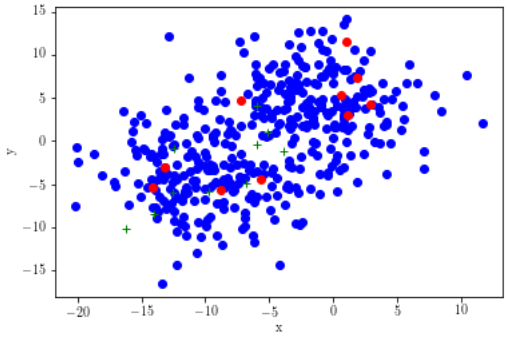


RKHS

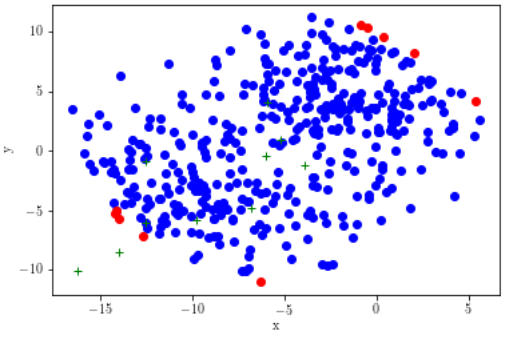


RKHS · variance

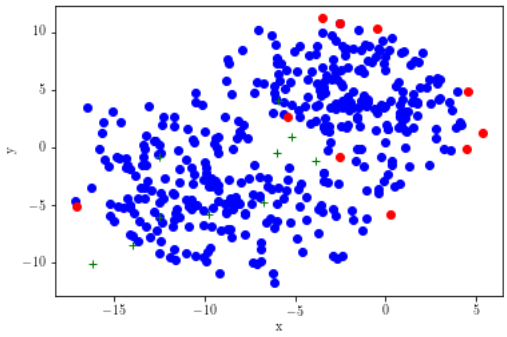
Figure 2: Blue points - unlabeled data given to the model to choose next point from. Red points - 10 points that were chosen. Training set contains 50 points.



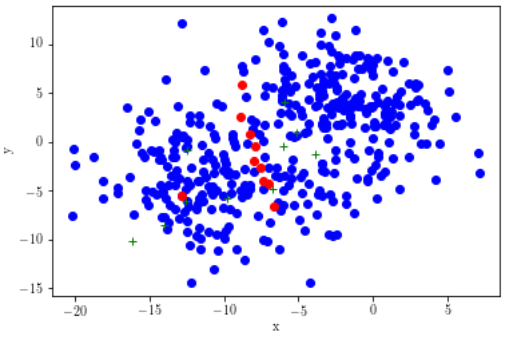
Random



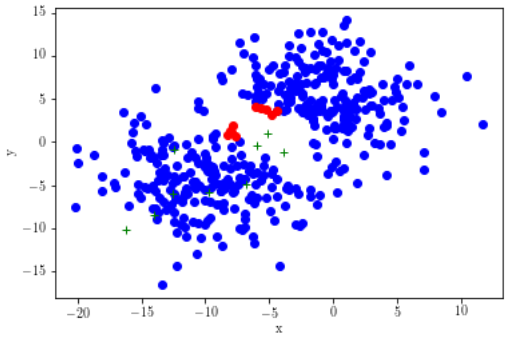
Max-variance



2-norm in formula

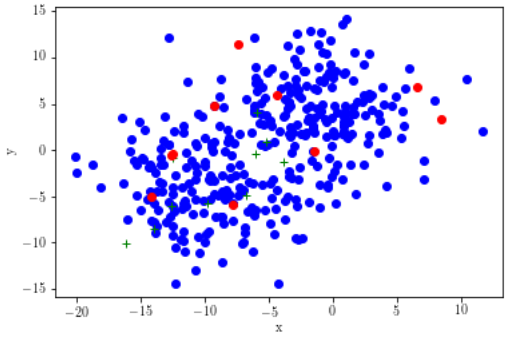


RKHS

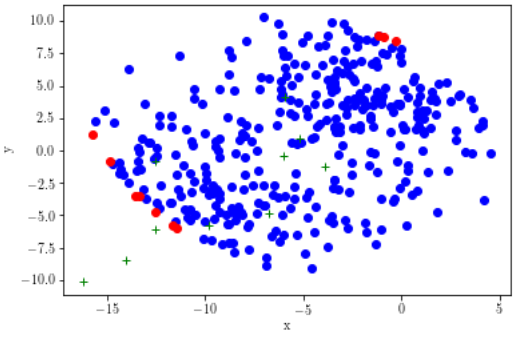


RKHS · variance

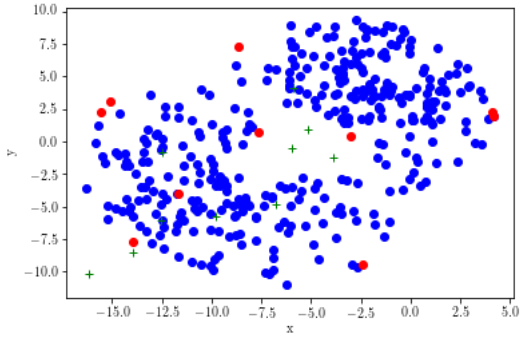
Figure 3: Blue points - unlabeled data given to the model to choose next point from. Red points - 10 points that were chosen. Training set contains 100 points.



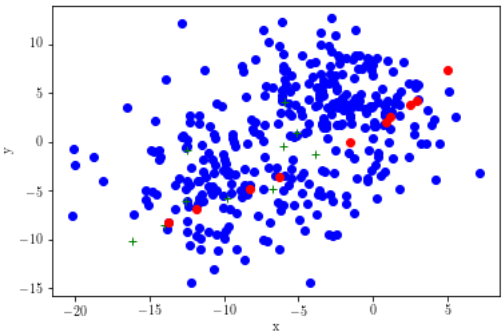
Random



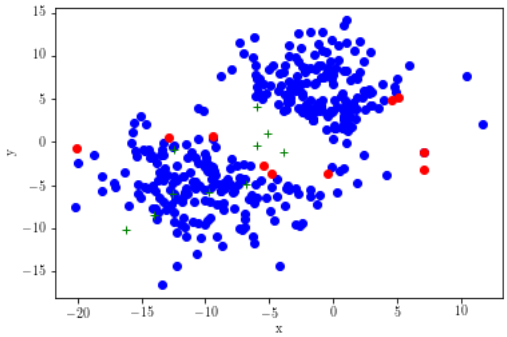
Max-variance



2-norm in formula



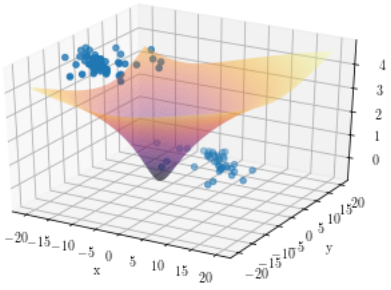
RKHS



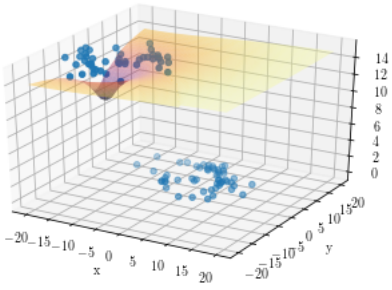
RKHS · variance

Figure 4: Blue points - unlabeled data given to the model to choose next point from. Red points - 10 points that were chosen. Training set contains 150 points.

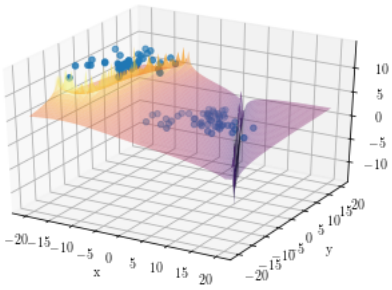
Next we took the same dataset and plot 3-d surfaces of score-functions and also tried to project them onto a plane using contour plots. The results are presented on figures 5 and 6. Training dataset is containing 100 points.



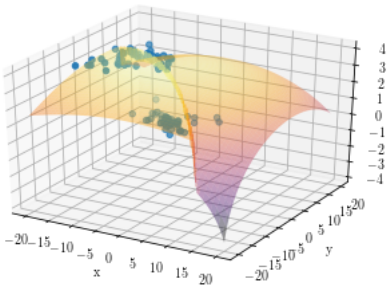
Max-variance



2-norm in formula

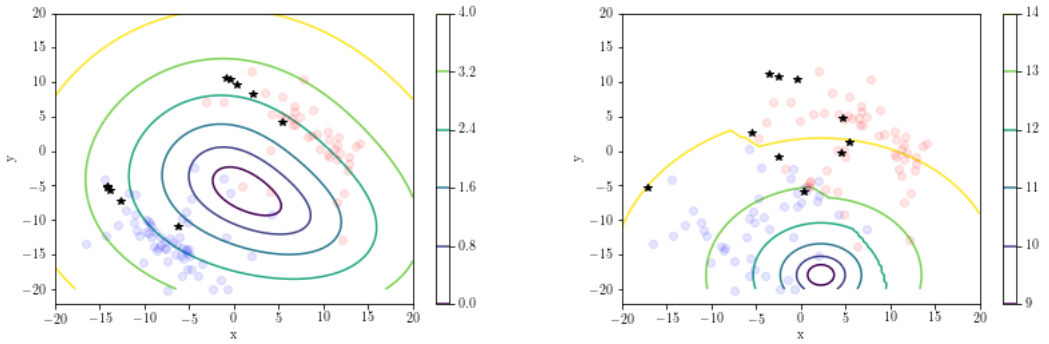


RKHS



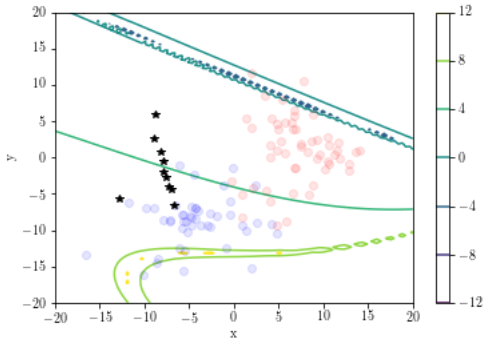
RKHS · variance

Figure 5: Blue points - training dataset. Surface is drawn on the grid 100x100 points in total. The logarithmic scale is selected along the third axis.

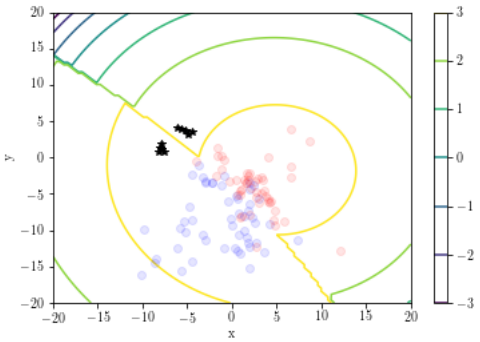


Max-variance

2-norm in formula



RKHS



RKHS · variance

Figure 6: Blue points - training examples from the first class. Red points - from the second class. Black stars - recently chosen points. Contour is drawn on the grid 100x100 points in total.

On the figure 7 accuracies depending on the size of training dataset and method are shown.

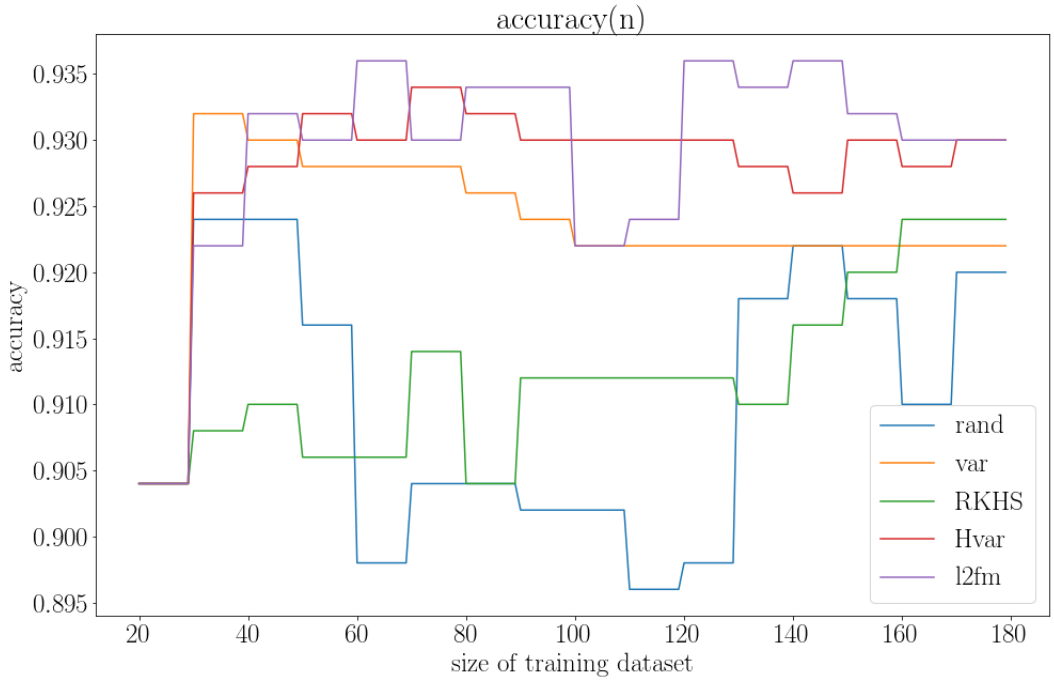
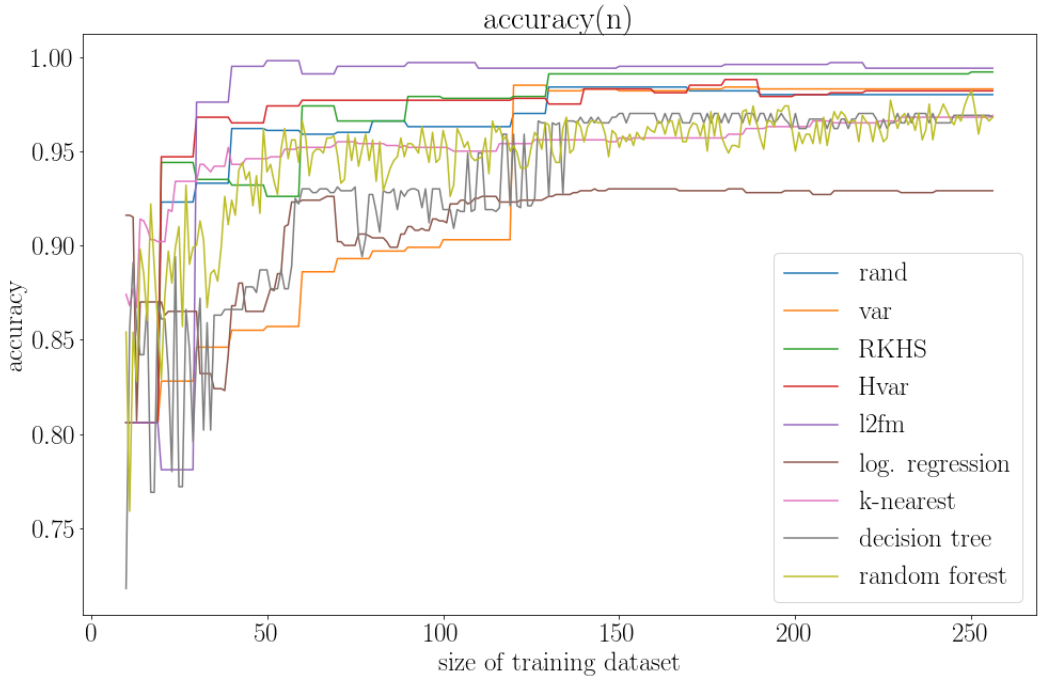


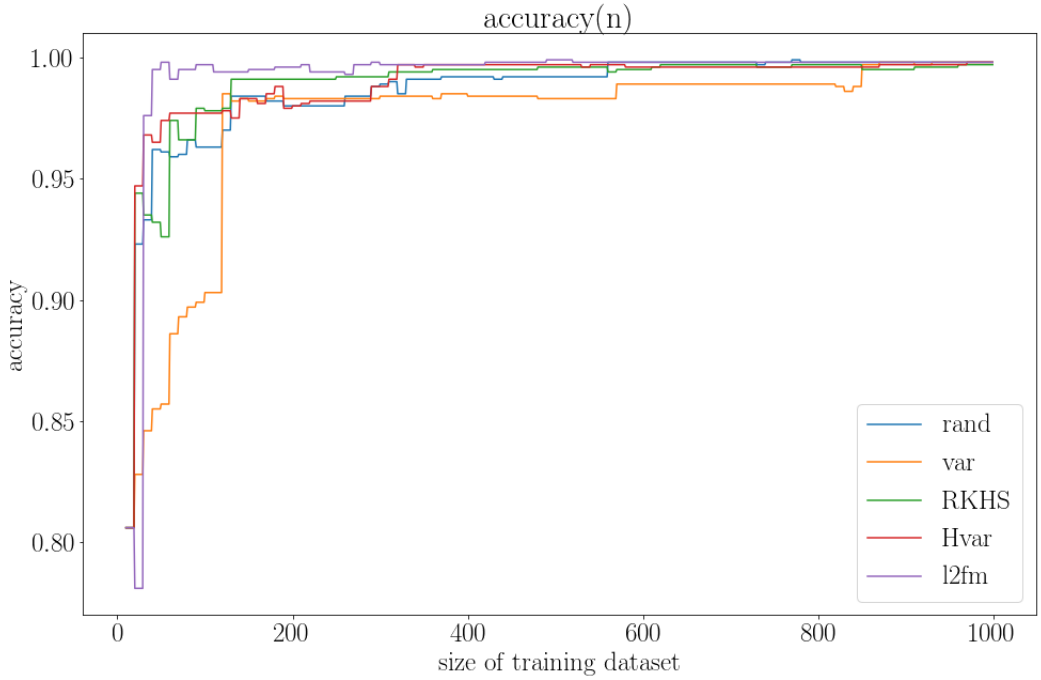
Figure 7: Accuracy depending on th size of training dataset. In addition to active learning criteria logistic regression and kNN-classifier were compared.

4.2 Skin

Skin dataset is collected by randomly sampling B,G,R values from face images of various age groups (young, middle, and old), race groups (white, black, and asian), and genders obtained from FERET database and PAL database. Total learning sample size is 245057; out of which 50859 is the skin samples and 194198 is non-skin samples.
We used 2000 random samples from the dataset. 1000 went to the test dataset and training dataset changed from 10 to 1000 points. We used it to compare not only different score-functions, but also compare active-learning approach with some traditional methods: logistic regression, k nearest neighbours, decision tree and random forest.
On the figure 8 you can see how accuracy of the certain method depends on the size of training dataset. Traditional methods shared train dataset with random score.



Size of train datset changes from 10 to 250 points.



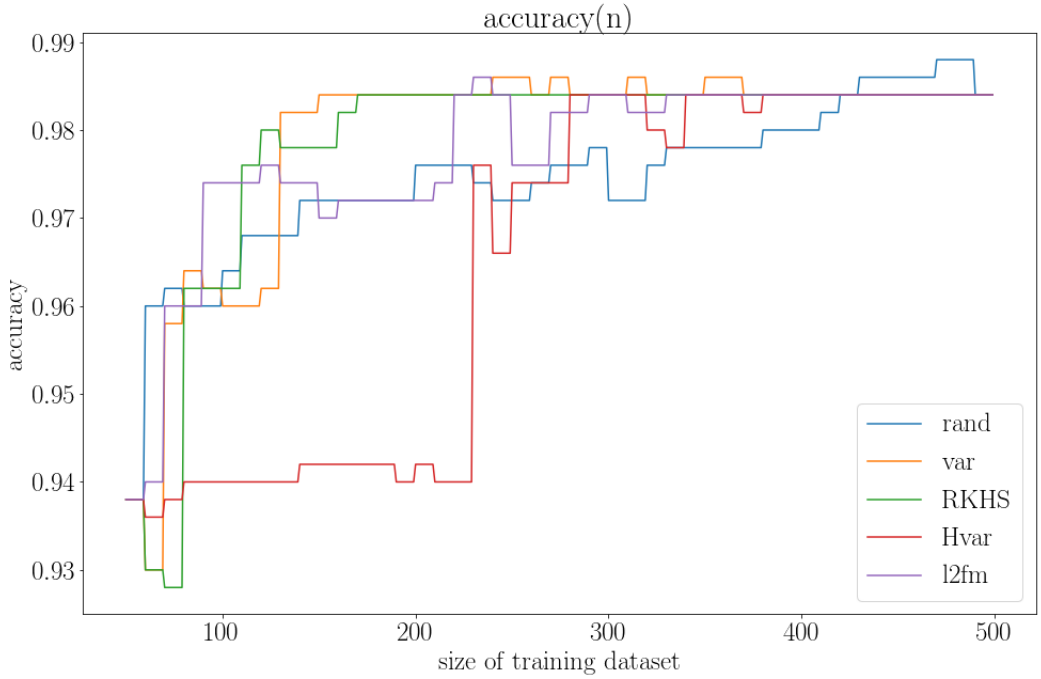
Size of train datset changes from 10 to 1000 points.

Figure 8: Compartion of different approaches to classification task on skin dataset.

4.3 HTRU2

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South).

Here the legitimate pulsar examples are a minority positive class, and spurious examples the majority negative class. At present multi-class labels are unavailable, given the costs associated with data annotation.
The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators.
We used total 1000 points from the dataset, test dataset containd 500 points, and the training dataset changed from 10 to 500 points.



Size of train datset changes from 10 to 500 points.

Figure 9: Compartion of different approaches to classification task.

4.4 Time comparsion

				Score Times				
Dataset Name	Dimensionality	Start train size	End train size	rand	mvar	RKHS	Hvar	l2fm
TwoDim	2	10	1000	544	501	488	487	512
HTRU2	8	10	500	33.23	33.32	33.98	34.92	37.43
Skin	3	10	1000	696	668	627	630	633

4.5 Conclusions

From the experiments we certainly can say that choosing new point in active learning process by the maximum of variance leads to worse quality. Other methods perfom better or at least as well as random sampling.

References

[1] Burnaev E., Panov M. (2015) Adaptive Design of Experiments Based on Gaussian Processes. In: Gammerman A., Vovk V., Papadopoulos H. (eds) Statistical Learning and Data Sciences. SLDS 2015. Lecture Notes in Computer Science, vol 9047. Springer, Cham

[2] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X

[3] Mina Karzand, Robert D. Nowak: Active Learning in the Overparameterized and Interpolating Regime arXiv preprint arXiv:1905.12782,2019

[4] Burr Settles. Active Learning Literature Survey. 2010.