Gaussian Process in Active Learning for Classification

Daria Kotova, Maxim Panov

November 28, 2019

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:

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

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
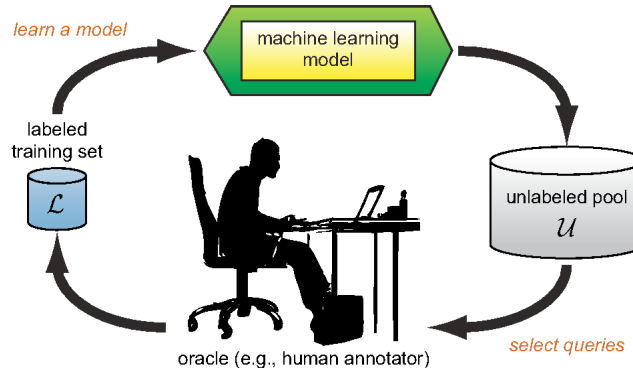

Figure 1: Illustration of active learning algorithm using pool-based method.

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):

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

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:

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

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

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

$$\|f_t^u(v)\| = \vec{y}^T K^{-1}\vec{y} + \vec{y}^T \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}}\vec{y} - 2t\vec{y}^T K^{-1}\frac{\vec{a}}{b} - 2t\vec{y}^T \frac{K^{-1}\vec{a}\vec{a}^T K^{-1}}{b - \vec{a}^T K^{-1}\vec{a}}\frac{\vec{a}}{b} + \frac{1}{b - \vec{a}^T K^{-1}\vec{a}}$$

$$= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a}\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} - 2t\frac{\vec{y}^T K^{-1}\vec{a}}{b}\left(1 + \frac{\vec{a}^T K^{-1}\vec{a}}{b - \vec{a}^T K^{-1}\vec{a}}\right) + \frac{t^2}{b - \vec{a}^T K^{-1}\vec{a}}$$

$$= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a}\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} - \frac{2t\left(\vec{y}K^{-1}\vec{a}\right)}{b - \vec{a}^T K^{-1}\vec{a}} + \frac{t^2}{b - \vec{a}^T K^{-1}\vec{a}}$$

$$= \vec{y}^T K^{-1}\vec{y} + \frac{\left(\vec{y}^T K^{-1}\vec{a} - t\right)^2}{b - \vec{a}^T K^{-1}\vec{a}} = \vec{y}^T K^{-1}\vec{y} + \frac{(f(u) - y(u))^2}{b - \vec{a}^T K^{-1}\vec{a}}.$$

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 $f_t^u(v) - f(v)$:

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

Finally, we get:

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

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.