Project Proposal for ”Super-Samples from Kernel Herding” Paper

Georgii Novikov

1 Introduction

Herding algorithm is a greedy algorithm in the form

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} < w_t, \phi(x) >$$
$$w_{t+1} = w_t + \mathbb{E}_p \, \phi(x) - \phi(x_{t+1}),$$

where $\mathcal{X}$ is some set, $\phi(x) : \mathcal{X} \to \mathcal{H}$ is a given function to some Hilbert space $\mathcal{H}$ (with scalar product $< \cdot, \cdot >$) (feature map in the terminology of the original paper) and $p(x)$ is some given distribution over $\mathcal{X}$. Authors reason, that this greedy algorithm solves (greedily) the problem of finding the subset of $\mathcal{X}$ that best approximates value $\mu_p = \mathbb{E}_p \, \phi(X)$ and it can be proven that this sampling method decreases error

$$\varepsilon_T = \mu_p - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \tag{1}$$

with a rate of $\mathcal{O}(T^{-1})$ (which is significantly better than for a subset of i.i.d samples, which has a rate of convergence of $\mathcal{O}(T^{-\frac{1}{2}})$; or even slower for MCMC samples due to the correlation of samples).

On the first sight, the problem, that this method is trying to solve, application area and the method as a whole is not clear at all: to apply it, we should already have a given set of points (so it cannot work with in some sense intractable probability distributions like MCMC), we have to be able to take expectations over p (which in the presence of the fact that the set $\mathcal{X}$ is assumed to be given and finite is not a problem, but still) and we have to know in advance the feature map $\phi(x)$. But if we would perceive this algorithm not like a sampling procedure, but like the greedy optimization procedure of (1), then everything becomes more clear: it can be applied to finding the Coresets [2] (where we want to find a small dataset subset on which we can calculate the lower bound of loss function for some prediction model family), dataset reduction (similar to Coreset, we want to find a small dataset subset on which the trained prediction model would be of the same quality as trained on the whole dataset), model compression [1] (approximate big model with smaller one), ensemble subsampling (when we want to remove unnecessary ensemble entities but preserve ensemble quality) and many others (but it is still not clear for me, how the performance of this procedure is related to the other methods of optimization)

2 Kernel Trick

Paper of interest is dedicated to the kernel trick for the herding algorithm. The initial problem was that in order to apply it, we had to have a finite-dimension Hilbert space

$\mathcal{H}$, because we keep the current state of approximated value w_t. If we substitute this value with its explicit form of $w_T = w_0 + \sum_{t=1}^{T} \phi(x_t)$ and replace explicit function $\phi(x_t)$ with an implicit in form of a kernel $k(x, x') =< \phi(x), \phi(x') >$, then new sample would be given with the following formula:

$$x_{T+1} = \arg\max_{x \in \mathcal{X}} \mathbb{E}_{x' \sim p}\, k(x, x') - \frac{1}{T+1} \sum_{t=1}^{T} k(x, x_t). \tag{2}$$

Here, under different assumptions, we can calculate $\mathbb{E}_{x' \sim p}\, k(x, x')$ explicitly or estimate over all set $\mathcal{X}$.

3 Experiments

Basically, that's it. Plug (2) into the following tests and everything would be fine:

1. Mixture of Gaussians for different kernels with explicit expectation in (2) (matching true distribution).
2. Mixture of Gaussians, but with expectation estimated over 10^5 i.i.d samples (matching empirical distribution).
3. Find supersamples in the set of logistic regressions, sampled with MCMC from the posterior distribution for the spambase dataset.

References

[1] Cristian Bucil, Rich Caruana, and Alexandru Niculescu-Mizil. *Model Compression*. Tech. rep. 2006.

[2] Alexander Munteanu et al. “On Coresets for Logistic Regression”. In: (May 2018). arXiv: 1805.08571. URL: https://arxiv.org/abs/1805.08571.