Sub-sampling Schemes for Kernel Methods

Abstract: Kernel methods are prevalent tools to address nonlinear regression and classification problems. The exact solution of these methods has the form $f(x) = \sum_{i=1}^{n} a_i K(x, x_i)$ with n basis functions. However, one significant problem is that the computational cost scales as $O(n^3)$. In order to reduce the complexity, this project focuses on using a sub-model $f(x) = \sum_{i \in S} a_i K(x, x_i)$ to approximate the exact solution, by sub-sampling a subset *S* of training data to form the solution. Novel strategies for sub-sampling will be investigated, and the performance will be evaluated on three kernel methods: Gaussian Processes (GP), Kernel Support Vector Machine (Kernel SVM), and Kernel Logistic Regression (KLR).

Objectives and Approach: the goal of this project is to propose a novel sub-sampling method for model reduction of kernel methods, and empirically investigate its performance on large-scale data sets. We are interested in researching by using sub-sampling method how many samples are large enough and how much training time could be reduced to achieve the same prediction error compared to using full data sets. One promising approach is to cluster the training set and selects only the cluster centers to form the basis functions for the sub-model, i.e. $f(x) = \sum_{i \in S} a_i K(x, x_i), i \in \{cluster\ centers\}$ and perform stochastic gradient descent to find the minimum solution $\vec{a}$, for example for the following KLR cost function, (λ is the regularization constant chosen by cross validation):

$$H(\vec{a}) = -\sum_{i=1}^{n} [y_i f(x_i) - ln(1 + exp(f(x_i)))] + \frac{\lambda}{2} \|f(x)\|_{\mathcal{H}\mathcal{K}}^2$$

To assess the performance of sub-sampling methods, we choose squared loss and 0-1 loss for GP, hinge loss for Kernel SVM and negative log-likelihood loss for KLR. The solution consisting of full training examples and the one using a smaller, uniformly sampled, subset will be used as two baselines in comparison with the proposed sub-sampling schemes. The major trade-offs behind the proposed sub-sampling approaches will also be explored.

Work plan and implementation:

1. Literature review of GP, Kernel SVM and KLR and theoretical study of their complexities.
2. Design and implement a kernel-method experimental framework in Python. Set up two baselines by the performance of models using full examples and a uniformly sampled subset.
3. Integrate the proposed sub-sampling schemes to the experimental framework.
4. Compare the performance of the proposed methods to the baselines in step 2 and summarize the findings.
5. Documentation and dissemination of the results.

Deliverables:

1. Concise, results reproducible and well-commented codes in Python.
2. A report describing the conducted work in full details, e.g. ideas, theoretical analysis, experimental results and discoveries.

3. All documents, including the report, the codes and the final presentation have to be saved on a CD-ROM and handed over to the supervisors at the latest at the final presentation.

Prof. Andreas Krause	Mr. Oliver Bachem	
Signature:	Signature:	Signature