Bayesian Multi–task Learning for Function Estimation with Dirichlet Process Priors

Marcel Hermkes, Nicolas Kuehn and Carsten Riggelsen
Institute of Earth and Environmental Science
University of Potsdam
Karl–Liebknecht Str. 24/25, 14476 Golm–Potsdam, Germany
{hermkes,nico,riggelsen}@geo.uni-potsdam.de

In this study we consider the problem of multi–task learning (MTL), strictly speaking, learning multiple related predictive functions, for which the assumption is that the training data set for each task is not statistically identically distributed, but that similar tasks share some information. In Probabilistic Seismic Hazard Analysis, for example, seismic ground motion data, induced by earthquakes, are collected at different geographical regions. Instead of building ground motion models (predictive functions) for each region individually, it is preferable to share information across the tasks to increase the overall prediction performance.

One of the most important methods to share information correlation between tasks is Hierarchical Bayesian modeling, where parameters of the task–specific models are coupled by a common prior. As a result of learning the parameters of the model and the hyperparameters of the common prior jointly, the function estimation of a specific task is affected by its own training data and by data from the other task related through the coupled prior. Generally, the common prior is specified in a parametric form with unknown hyperparameters. Such a parameterization has two drawbacks: (i) Individual tasks can be of high complexity and therefore the underlying common prior with an appropriate functional form may be difficult to decide upon. (ii) The relationship between all tasks are treated equally, it is however desirable that only similar tasks, which should be automatically identified, share information to permit negative transfer.

To deal with these issues we propose a nonparametric hierarchical Bayesian model where the common prior is drawn from a Dirichlet Process (DP). Such a nonparametric prior has the ability to fit the model well with respect to the data without restriction about the functional form of the prior distribution. Furthermore, the employed DP prior induces a partition of tasks, so that similar tasks within each cluster share the same parameterization. First of all, we present a linear regression model, for which the weights of the covariates and the model variance are drawn from a DP prior. As base distribution for the DP we have chosen a normal inverse–Gamma prior which is the natural conjugate prior to the normal likelihood of the applied regression model. In addition, we extend this model by replacing the linear regression model by Gaussian Processes. The resulting models can be seen as a DP Mixture (DPM) of linear regression functions, respectively DPM of Gaussian Processes. By choosing conjugate priors, the base distribution can be analytically marginalized, but the sum over all latent partitions makes exact Bayesian inference intractable. Instead of using MCMC sampling machinery which may be slow to convergence, we apply the Bayesian Hierarchical Clustering (BHC) algorithm (Heller and Ghahramani, Proceedings of ICML'05) to make approximative inference. The BHC constructs a lower bound of the DPM marginal likelihood, which approximates the sum over the latent partitions by summing over the exponential number of partitions of tree–consistent partitions, induced by the agglomerative clustering.

So far, scant attention was paid to the prediction of novel tasks, for which no training data are available. Assuming the existence of task dependent features, we can deal with this problem by extending our model with an additional Gaussian component over these task features for each specific task, which are also coupled by the DP prior. This component is considered as gating function, that determines the responsibility for each task with respect to the novel task.

The experiments were conducted on basis of two real world problems: (i) prediction of ground motion intensity parameters (Allen and Wald, USGS Open–File Report 2009–1047) and (ii) exam score prediction (`http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-support/datasets.shtml`). For each dataset we also consider the performance of a single task learning method (training a separate model for each task) and a complete pooling approach (train a model on the complete data) as baseline methods. The results show improved prediction performance of the DPM models compared to these baseline methods.