STAT 771 - Project Summary

Stewart Kerr, Prof. Menggang Yu, Prof. Guanhua Chen
Department of Biostatistics and Medical Informatics

Mixture proportion estimation (MPE) is the problem of estimating the weight of a component distribution in a mixture given samples from the mixture and component. Solving this problem is an important step in "weakly supervised" learning tasks in which one has access to only positively labelled data or there is noise in the labels in the training set. Specifically, we are interested in MPE to assess the similarity between two different populations of hospital patients. For example, we wish to use data observed from patients in Iowa to improve models learned using data from Wisconsin patients. That is, Iowa is defined as our **source** population while Wisconsin is our **target** population. Among several methods proposed to solve MPE problems, we adapt methods proposed by Ramaswamy et. al via kernel embedding of distributions. This approach utilizes an efficient algorithm for MPE along with guaranteed convergence to estimate the true proportion under certain less restrictive conditions. Using this method, we can obtain a similarity estimate, $\hat{\kappa}$, of the proportion of the source population that matches the target population.

We wish to extend the mixture proportion estimate to identify the $\hat{\kappa}$ proportion of observations from the source population which most closely match the target. Successful identification of these observations will allow us to combine a subset of the source population with the target population for subsequent model learning. To identify these observations, we propose using SVM classification methods to estimate a separating hyperplane to discriminate the source and target populations. Specifically, we will build an SVM classifier using the same kernel used for MPE estimation above. SVM is a desirable approach because distance from each observation to the separating hyperplane provides a direct measure of how similar an observation from the target population is to samples in the source population.

Existing code published by Ramaswamy et. al for MPE can be adapted for our use. We plan to write additional code to perform SVM classification of the source and target populations and identify which observations from the source population most closely match the target. Ideally, we will test our program on simulated data before applying it to EHR and insurance claims data obtained from both UW-Health and University of Iowa Health. We estimate approximately 2-3 weeks to gain additional understanding of the task, followed by 3 weeks of coding, followed by 2-3 weeks of applying the method and writing up results.