Thomas G. Stewart

tgs@email.unc.edu • www.bios.unc.edu/∼tgs

Research Statement

1. Current Research

Support Vector Machines with Missing Data
Statistical learning is a popular family of data analysis methods particularly suited for high dimensional settings. Statistical learning methods have been successfully employed in biomedical research, the social sciences, public safety applications, and most data dependent areas of research. One goal of statistical learning methods is to construct rules which predict an outcome y from a potentially large set of predictors $\mathbf{x}$, for example, predicting treatment response from a set of pre-treatment biomarkers. The support vector machine (SVM) is a statistical learning method introduced in Boser et al. (1992). SVMs have been successfully employed in both classification and regression tasks, and the method is particularly useful in computer vision applications. It is a basis expansion method which provides the user with considerable modeling flexibility.

Missing data are ubiquitous. Despite continuing advances in data collection, missing data are likely to remain a permanent feature of statistical analysis. My dissertation research provides principled methods for constructing SVM classifiers when the training set includes missing values. These methods are important to users of support vector machines because missing data is common-place and alternative approaches are ad-hoc or computationally intractable.

AWSVM. This method is an EM-type solution for missing values in the covariates. The basic idea is to replace the empirical risk component of the SVM objection function with its expectation based on a specific conditional EM distribution. Because the SVM classifier is a parameter of the conditional EM distribution, one can alternate between calculating the conditional expectation and computing an SVM solution based on the new expectation. I showed that the resulting classifier is a Bayes classifier when the distribution of the covariates is properly specified.

DRSVM. Because AWSVM has a number of computational challenges and has limited applicability to high dimensional settings, I developed DRSVM. This second method, also for missing values in the covariates, is similar to doubly robust type estimators common in estimating equation inference. The weighted estimating equation framework for missing data avoids many of the drawbacks of the EM-type method. Like linear regression, the weighted estimating equation solution for SVMs reweights the empirical risk component of the SVM objective function and requires (a) estimation of selection probabilities and (b) a surrogate loss function usually computed via imputation. As such, it does not require a covariate distribution assumption or the estimation of nuisance parameters. I showed that the DRSVM is asymptotically a Bayes classifier when either the selection probabilities or the surrogate loss are specified correctly. As such, the DRSVM is a doubly robust missing data solution for SVMs.

2. Future Research Plans

Causal Inference and Statistical Learning
There are a number of research questions that I am positioned to pursue because of the combination of (a) my dissertation research and (b) several collaborative projects involving causal inference and observational, longitudinal patient data. A key issue in observational, comparative effectiveness studies is to account for potential confounding due to treatment choice. Current practices to adjust for covariates related to treatment choice include doubly robust type estimators, stratification, or matching. Each of these methods can and often use propensity scores, and researchers have recently implemented propensity scores constructed from statistical learning methods (Austin, 2011). Such research is an important first step towards implementing statistical learning ideas as solutions to traditional statistical issues like confounding. There is still ample work to be done in this area. For example, recent work with covariate balancing propensity scores (Imai and Ratkovic, 2014) improves on earlier parametric propensity score methods, and extending such methods with statistical learning methods is potentially valuable to causal inference and comparative effectiveness research.

I also propose research which considers how traditional issues like confounding can affect the usefulness of modern statistical learning methods. Because large, observational databases are and will continue to be the focus of modern research, statistical learning methods which account for potential confounding are, I propose, valuable. My work in missing data with SVMs has already provided a basis for future research in this area because of the close connection between missing data and causal inference methods. For example, reweighted SVMs like DRSVM can be adapted to correct for potential confounding from treatment choice in ways similar to how the method adjusts for "confounding" from missing data.

Statistical Learning and Missing Data
There are a number of research questions related to SVMs and missing data that deserve further consideration. The methods described above were developed in the context of covariates Missing at Random (MAR). A natural next step is to consider situations when the covariates are Not Missing at Random (NMAR). In parametric settings, one way to analyze NMAR data is to explicitly model the missingness mechanism. The outcome/prediction model and the missingness model are estimated jointly such that if both models are specified correctly, then the estimated parameters are unbiased. A similar approach with SVMs is a reasonable starting point for this research question. As a statistical learning approach, the missingness model could be non-parametric.

In addition to SVM methods for NMAR missing data, there is also the more general question of whether one can even determine whether missing data are MAR or NMAR. In parametric settings, some procedures exist to test this question. One area of future research is to consider these procedures in high dimensional settings when most of the covariates are not predictive of the outcome.

References

Imai K, Ratkovic M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, Pittsburgh, PA, 1992.