Summary 24: Linear-Time Estimators for Propensity Scores

Susan Cherry

Propensity scoring is used to answer "what if" questions. For example, if there is a new drug and it is tested on a random set of patients, we want to know what would happen if it had been administered to everyone. This paper focuses on propensity scoring algorithms with large control and treatment groups. The paper begins by describing Propensity Scoring and Covariate Shift and Correction. p is referred to as the treatment distribution and q as the control. The assumption of Relative Density is made, which means that the control group can't contain any significant component of patients that have very different covariates than the treatment group. They then briefly outline propensity scoring and covariate shift correction, in which we assume that $p(y|x) = q(y|x)$. If the distribution of covariates for the test group is different than the training, then the goal is to find a risk minimizer that minimizes respected risk with respect to $p(y|x) = q(x)$. The next section discusses Logistic Regression for propensity score estimation. Then they introduce convex duality and operator mean matching. Specifically, they discuss mean matching, kernel mean matching, and entropy regularization.

Next, online algorithms for distribution matching are discussed. This is a constrained quadratic problem or a general convex problem. Scaling (as seen in MEMM) can be insufficient for large scale problems so this section presents online algorithms. The assumption that $\phi(x)$ is in the form of a feature vector is made. Next, the online algorithm for kernel mean matching is presented. the idea is to solve the problem in Theorem 4 (this is described in detail on page 96) instead of the original convex problem. A linear time algorithm for convergence of the KMM algorithm is established. Next, Markov Chain Monte Carlo for Maximum Entropy Mean Matching is presented. The disadvantage of the algorithm is that it requires a large number of passes through the data. MCMC estimation of the gradient is used to overcome this problem. Next, the Moving Average Estimation for Log Partition Functions is introduced. Instead of drawing from the gradient, Metropolis-Hastings is used to compute the running average of R_t for $g(\theta)$, which is used as an approximation for the denominator.

Next, we move to experiments. The online kernel mean matching algorithms are applied to a small UCI dataset to ensure they produce the correct propensity score estimates. Then the online algorithms are used on real world data. They find that estimates are close to true propensity scores, especially for small scores. Regularization reduces variance but increases bias, so the difference is larger for higher propensity scores. Next, they try to estimate the number of users who visit Yahoo each day. Their results show that the online algorithms work well and that propensity score-weighted estimates are all quite close to their true values.

In summary, this paper presents linear time online algorithms to estimate propensity scores and covariate shift correction coefficients. They find that their proposed algorithms are a significant improvement over previous work. The entropic algorithms perform slightly better than other moment matching methods, though the parameter setting is a bit more difficult. The authors conclude by recommending that the MCMC algorithms should probably be the first choice in practice.