Stratified Learning: A general-purpose method for learning under covariate shift with applications to observational cosmology

Roberto Trotta*[1,2], Maximilian Autenrieth[3], David Van Dyk[3], and David Stenning[4]

[1]Department of Physics [Imperial College London] – United Kingdom
[2]Scuola Internazionale Superiore di Studi Avanzati / International School for Advanced Studies – Italy
[3]Department of Mathematics [Imperial College London] – United Kingdom
[4]Department of Statistics and Acturial Science Simon Fraser university – Canada

Abstract

Supervised machine learning will be central in the analysis of upcoming large-scale sky surveys. However, selection bias for astronomical objects yields labelled training data that is not representative for the unlabelled target data distribution. This affects the predictive performance with unreliable target predictions.
We propose a novel, statistically principled and theoretically justified method to improve learning under such covariate shift conditions, based on propensity score stratification, a well-established methodology in causal inference. We show that the effects of covariate shift can be reduced or altogether eliminated by conditioning on propensity scores. In practice, this is achieved by fitting learners on subgroups ("strata") constructed by partitioning the data based on the estimated propensity scores, leading to balanced covariates and much-improved target prediction.
We demonstrate the effectiveness of our general-purpose method on contemporary research questions in observational cosmology, and on additional benchmark examples, matching or outperforming state-of-the-art importance weighting methods, widely studied in the covariate shift literature. We obtain the best reported AUC (0.958) on the updated "Supernovae photometric classification challenge" and improve upon existing conditional density estimation of galaxy redshift from Sloan Data Sky Survey (SDSS) data.
Slides: in PDF

Video: https://youtu.be/abUN2QDQ3dI

Keywords: Machine learning, covariate shift, selection effects, cosmology, supervised learning

*Speaker