@R_Trotta

StratLearn:

A general-purpose statistical method for improved learning under Covariate Shift

Max Autenrieth

Imperial
(Statistics Section)

Roberto Trotta

Imperial
&
International School for Advanced Study (SISSA)

David Stenning

Simon Fraser University

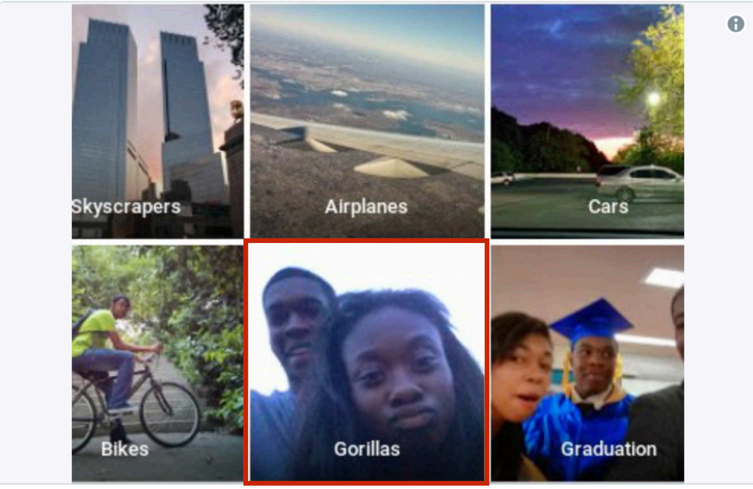
David van Dyk

Imperial
(Statistics Section)

Oct 20th 2021

The problem: machine learning classifiers trained on non-representative data generalize poorly.

2015

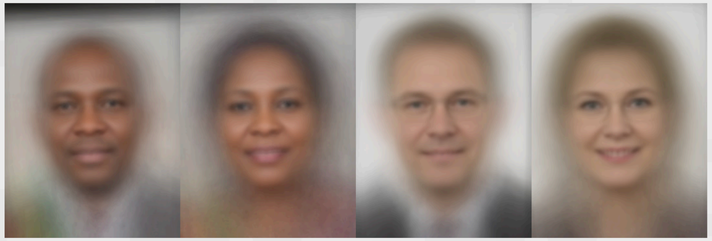


Jacky lives on @jalcine@playvicious.social now.
@jackyalcine

2019

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Gender Shades (MIT Media Lab, 2019)



Poor accuracy in facial recognition for dark skinned females

Cosmology

Incorrect classification of Type Ia vs non-Ia from photometric data leads to cosmological parameters systematic bias.

Non-representative spectroscopic training sample leads to incorrect photo-z estimation

2018

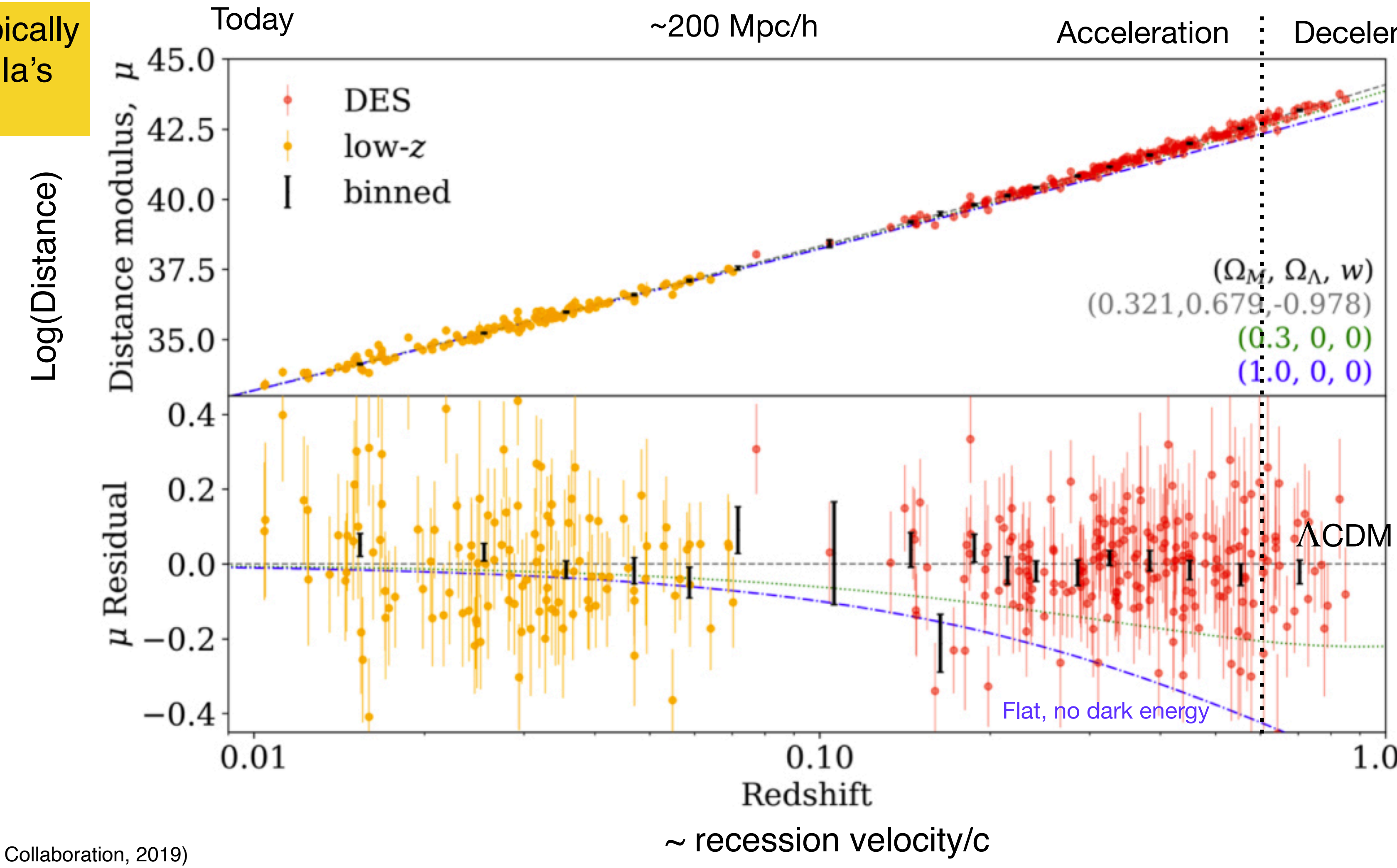
TOM SIMONITE BUSINESS 01.11.2018 07:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

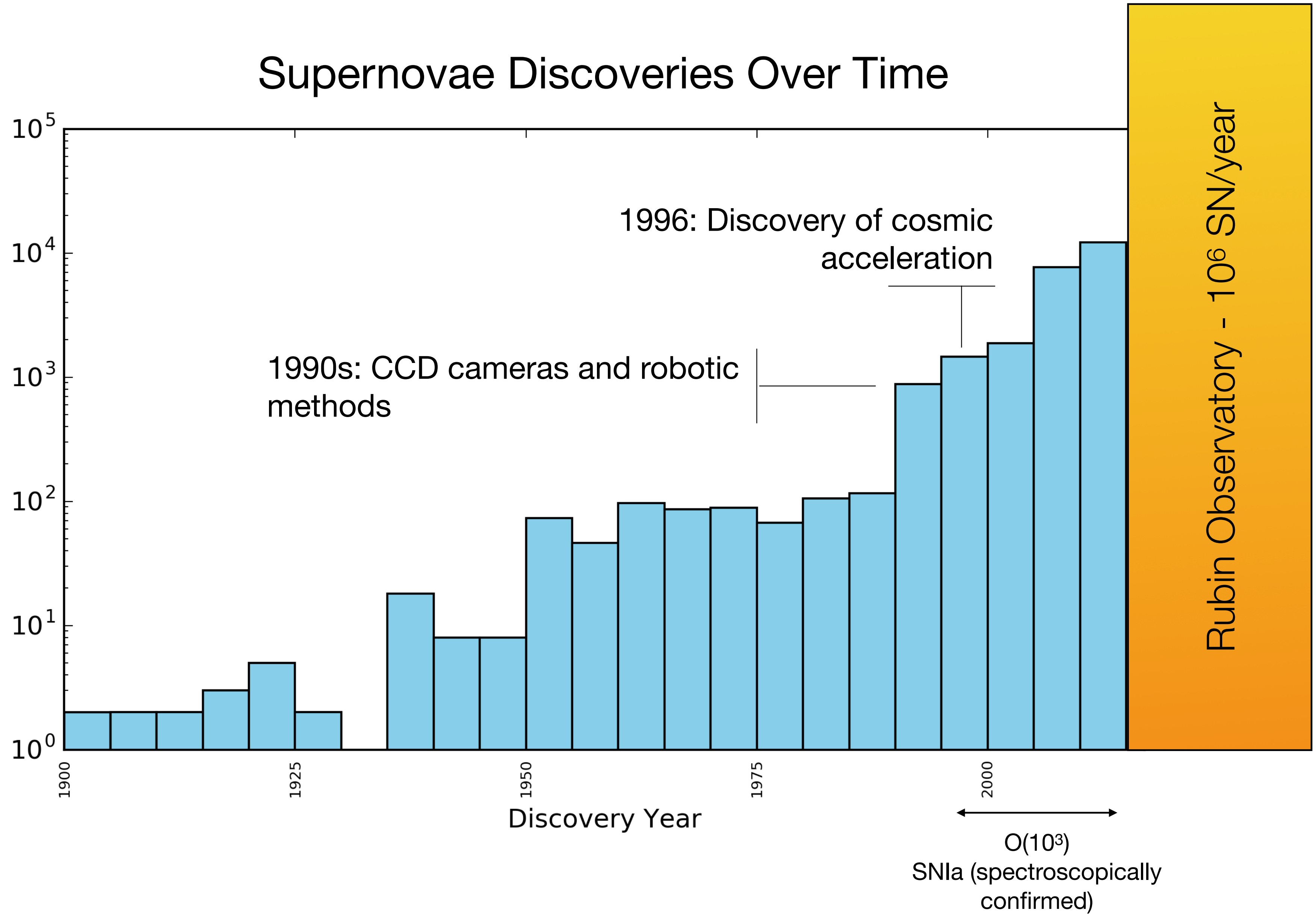
Distance-Redshift Relation Measurement

Spectroscopically confirmed Ia's only



~ 0.25 mag fainter than w/o dark energy

Abbott et al (DES Collaboration, 2019)



Supernovae Discoveries Over Time
10^5
10^4
10^3
10^2
10^1
10^0
1996: Discovery of cosmic acceleration
1990s: CCD cameras and robotic methods
Rubin Observatory - 10^6 SN/year
1900
1925
1950
1975
2000
Discovery Year
O(10^3) SNIa (spectroscopically confirmed)

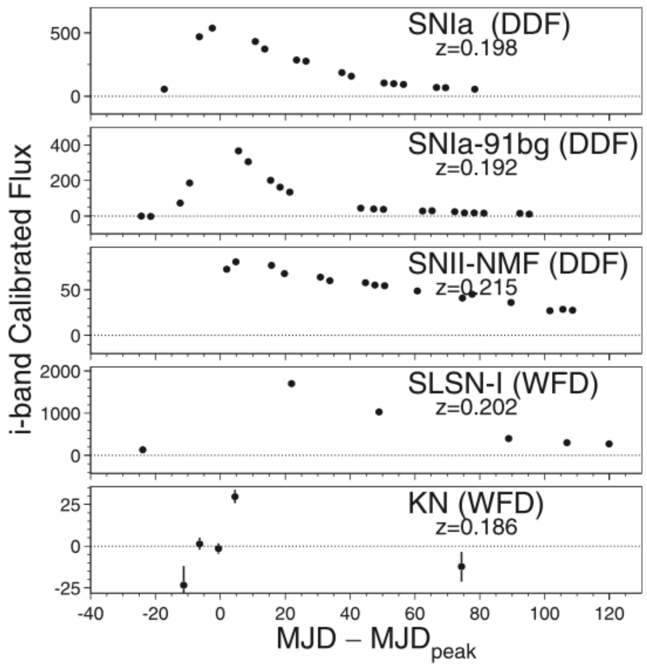
The Problem:

We want to classify Type Ia vs non-Ia **reliably** and **efficiently** from light-curve data alone.

BUT:

Spectroscopic training set is non-representative.

Simulated light-curves



Kessler et al (2019)

<u>Classification challenges:</u>

The Photometric LSST Astronomical Time-series Classification Challenge PLAsTiCC (Kessler et al, 2019)

Supernova Photometric Classification Challenge (Kessler et al, 2010)

Covariate Shift, or Biased Training Set

Light-curve data Type Ia or not

Given a feature space, X, and a label space, $Y\,(K > 1$ classes/dependent variables)

Spectroscopic training set

we have n_s labelled samples $\{x_i^s, y_i^s\}$ from the source domain

Photometric light-curve only

n_t unlabelled samples from the target domain, $\{x_i^t\}$.

Is it a Ia?

Task: predict $\{y_i^t\}$

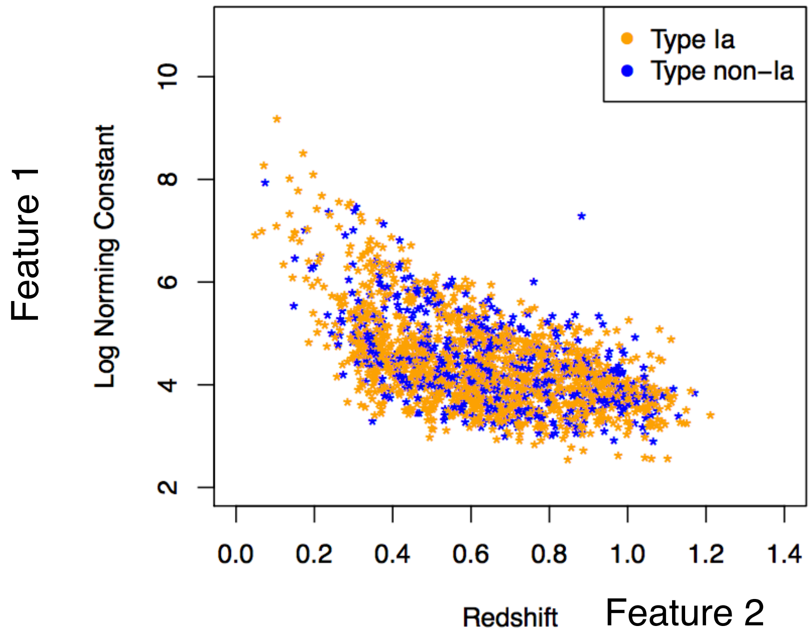
Covariate shift occurs when:

$$p_s(y\,|\,x) = p_t(y\,|\,x)$$

and $p_s(x) \neq p_t(x)$

I.e., the training set is non-representative of the test set.

Features: redshift & apparent mag
Label: Ia or non-Ia



Source domain (Coloured dots)

Target domain (Gray dots)

Revsbech, RT, van Dyk (2018)

Our Approach: Propensity Score Stratification

Work by **Max Autenrieth** (Stats PhD student), in collaboration with David van Dyk (Imperial) & David Stenning (Simon Fraser U.)

Improving on our previous work ("STACCATO"), Revsbech, RT, van Dyk (2018)

Propensity scores

$e(x_i)$ = probability for object i to be selected into the source domain:

$$e(x_i) \equiv P(s_i = 1 \mid x_i)$$

Idea (StratLearn):

subdivide ("stratify") target and source data in k subgroups according to quantiles of their propensity scores. Then supervised learning in each stratum ("stratified learner")

Propensity scores as balancing scores

Rosenbaum & Rubin (1983, 1984) show that, conditional on their propensity scores, the k subgroups ("strata") have approximately balanced covariate distribution, i.e.

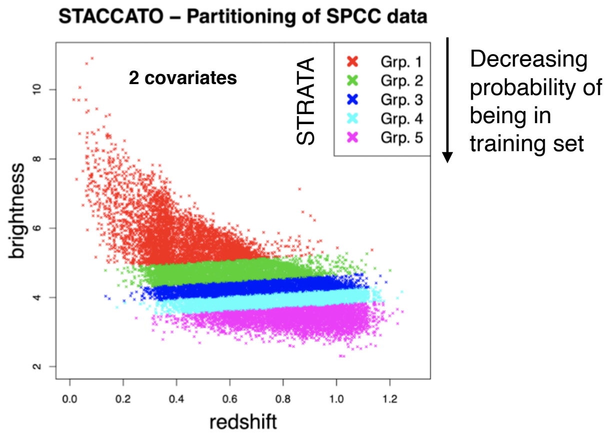
$$p_{s_j}(x) \approx p_{t_j}(x) \text{ for } j = 1, \ldots, k$$

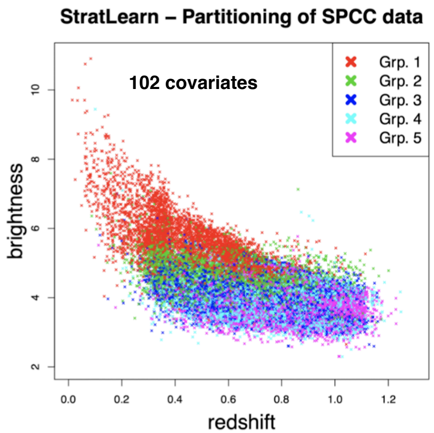
Since $p_s(y \mid x) = p_t(y \mid x)$, it follows that

$$p_{s_j}(x, y) \approx p_{t_j}(x, y) \text{ for } j = 1, \ldots, k$$

StratLearn on SNIa data

Propensity score partitioning of target domain (test data):





Conditional on the propensity scores (i.e., within each stratum), the source and target outcomes are approximately the same.

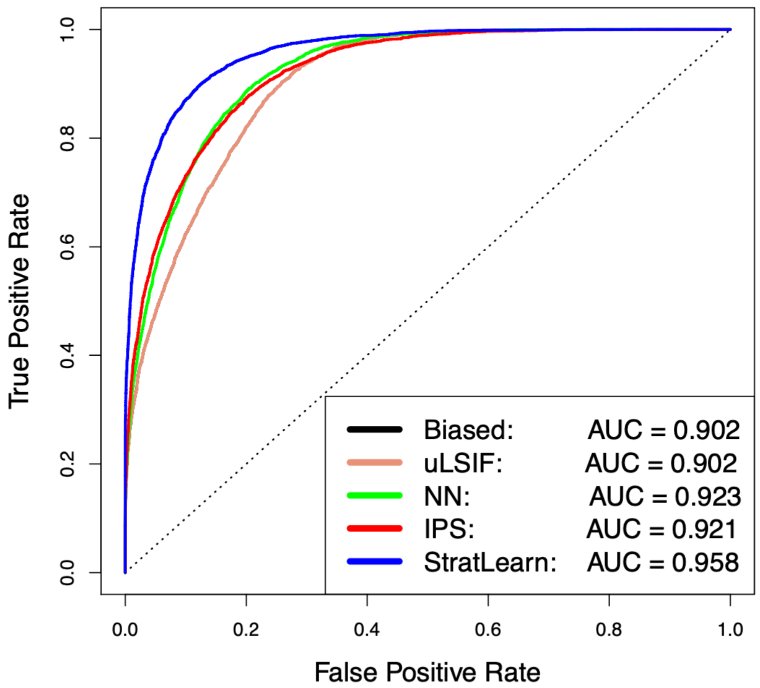
This means: inside each stratum, the imbalance has been redressed, i.e. source data are approximately representative

Important: the underlying theorem only valid if all potential confounding covariates (i.e., things the SNIa type could depend on) are included in the propensity score estimation!

Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa	
1	Source	958	518	0.54	👍 Balanced proportions
	Target	3306	1790	0.54	
2	Source	120	28	0.23	👍 Balanced proportions
	Target	4144	927	0.22	
3	Source	13	4	0.31	
	Target	4250	540	0.13	
4	Source	7	4	0.57	
	Target	4257	610	0.14	
5	Source	4	4	1	
	Target	4259	662	0.16	

Classification/Regression with StratLearn

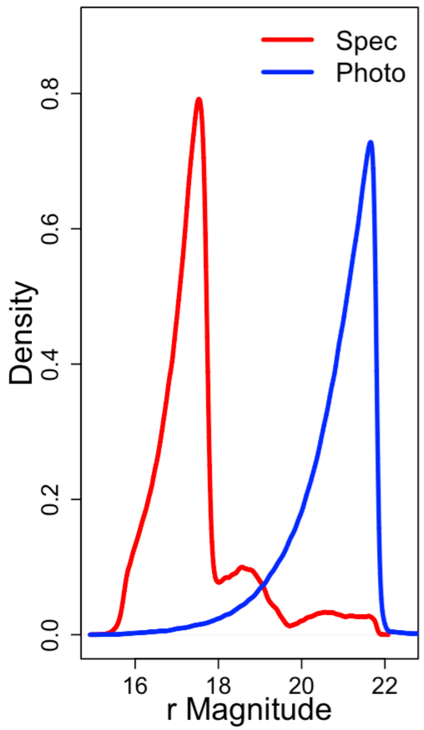
SNIa photometric classification (SPCC Challenge, v2)



StratLearn performance (AUC = 0.958) close to “gold standard” of unbiased training set (AUC=0.977) without any augmentation, beats all previous methods:

- Lochner et al (2016): AUC= 0.855
- Pasquet et al (2019): AUC=0.939
- Revsbech et al (“STACCATO”, 2018): AUC=0.94

Photo-z estimation



StratLearning outperforms previous methods for this problem.

Performance improvement is larger in the presence of high-D noisy covariates.

Note: AVOCADO (Boone, 2019), winner of the PLASTiCC challenge 2019, uses an extended version of STACCATO (incl. augmentation).

Conclusions

Imperial College London & SISSA DATA SCIENCE Machine Learning for the Natural Sciences

1. Covariate shift is an important and recurrent phenomenon in supervised learning. In dark energy research, it will affect the next generation of large SNIa data.

2. We propose a general approach (*StratLearn*) based on stratifying source and target domain according to propensity scores (= probability of an object to be included in the source domain).

3. Within strata, source and target domains are better balanced: StratLearn shows improved performance in regression and classification tasks compared to best-in-class alternatives.

Thanks to my collaborators: Max Autenrieth (PhD student), David van Dyk (Imperial), David Stenning (Simon Fraser U.). Paper here: https://arxiv.org/abs/2106.11211

Opportunities in (Data Science) x (Astro) at SISSA:

Currently open: Postdoc position (2+1 years)

Women and candidates from under-represented groups particularly encouraged!

Deadline: Nov 11th 2021

https://academicjobsonline.org/ajo/jobs/20085

Currently building a new data science group in Trieste, Italy

SISSA DATASCIENCE
Machine Learning for the Natural Sciences

datascience.sissa.it

References (Astro)

- E. A. Revsbech, R. Trotta, D. A. van Dyk, STACCATO: a novel solution to supernova photometric classification with biased training sets. *Monthly Notices of the Royal Astronomical Society* **473**, 3969-3986 (2018).
- R. Kessler *et al.*, Models and Simulations for the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific* **131**, 094501 (2019).1.
- Boone, K., Avocado: Photometric Classification of Astronomical Transients with Gaussian Process Augmentation. *The Astronomical Journal* **158**, 257 (2019).
- R. Kessler, et al (2010), Supernova Photometric Classification Challenge. ArXiv:1001.5210
- T. M. C. Abbott *et al.* (2019), First Cosmology Results using Type Ia Supernovae from the Dark Energy Survey: Constraints on Cosmological Parameters. *The Astrophysical Journal* **872**, L30.
- M. C. March, R. Trotta, P. Berkes, G. D. Starkman, P. M. Vaudrevange (2011), Improved constraints on cosmological parameters from Type Ia supernova data. *Monthly Notices of the Royal Astronomical Society* **418**, 2308-2329.
- S. R. Hinton *et al.* (2019), Steve: A Hierarchical Bayesian Model for Supernova Cosmology. *The Astrophysical Journal* **876**, 15.
- H. Shariff, X. Y. Jiao, R. Trotta, D. A. van Dyk (2016), BAHAMAS: New Analysis Of Type Ia Supernovae Reveals Inconsistencies With Standard Cosmology. *Astrophys. J.* **827**, 25.
- D. Rubin *et al.* (2015), UNITY: Confronting Supernova Cosmology's Statistical and Systematic Uncertainties in a Unified Bayesian Framework. *The Astrophysical Journal* **813**, 137 (2015).
- J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer, D. Poznanski (2012), Semi-supervised learning for photometric supernova classification. *Monthly Notices of the Royal Astronomical Society* **419**, 1121.

References (Stats)

- Shimoidara (2000), Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference 90,* 2, 227–244.
- Zadrozny (2004), Learning and evaluating classifiers under sample selection bias. In Proceedings of the 21st international conference on Machine learning. ACM, 114.
- Rosenbaum, P. R. and Rubin, D. B. (1983), The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1, 41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association 79,* 387, 516–524.
- Chen, X., Monfort, M., Liu, A. & Ziebart, B.D.. (2016). Robust Covariate Shift Regression. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, in PMLR 51:1270-1279