Supervised Learning with Biased Training Data and Applications to Supernova Type Ia Cosmology
Setting the scene:
Supernovae Type Ia and the accelerated expansion

The statistical problem:
Supervised learning in the presence of training set bias

Idea:
Propensity Score Stratification approach

“Stratified Learning” approach:
Performance in classification and regression problems

Conclusions
What is dark matter?

What makes the cosmos accelerate?

Is there life elsewhere?

5% Normal Matter

25% Dark Matter

70% Dark Energy
Big Bang

$t = 0$

End of the visible cosmos
$t = 380,000$ yrs
$z = 1,100$

Radiation (decelerates)

$t = 7$ Bn yrs
$z \sim 0.7$

Dark matter (decelerates)

$t = 13.7$ Bn yrs
$z = 0$

Dark energy (accelerates)

Inflation
$t = 10^{-32}$ s (exponential)

Big Bang Nucleosynthesis
$t = 3$ mins

SN Ia
Cosmic Acceleration

Abbott et al (DES Collaboration, 2019)

~ recession velocity/c

Today

~200 Mpc/h

Acceleration

Deceleration

Distance modulus, \( \mu \)

Distance

DES

low-z

binned

\( \Omega_M, \Omega_{\Lambda}, w \)

\((0.321, 0.679, -0.978)\)

\((0.3, 0, 0)\)

\((1.0, 0, 0)\)

\(\Lambda CDM\)

~ 0.25 mag fainter than w/o dark energy

Flat, no dark energy
Combination of multiple probes **is required** to determine the cosmological parameters to high precision.

Baryonic Acoustic Oscillations (BAO) and Cosmic Microwave Background (CMB) are standard rulers.

**Planck Collaboration (2015)**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega_m$</td>
<td>$0.316 \pm 0.009$</td>
</tr>
<tr>
<td>$H_0$</td>
<td>$67.8 \pm 0.9$ km/s/Mpc</td>
</tr>
<tr>
<td>$\Omega_\Lambda$</td>
<td>$0.684 \pm 0.009$</td>
</tr>
<tr>
<td>$</td>
<td>\Omega_k</td>
</tr>
<tr>
<td>Age</td>
<td>$13.796 \pm 0.029$ Gyr</td>
</tr>
</tbody>
</table>

How accurate are these results?

Supernovae Type Ia: Observations

DETECTION

SN1994D in NGC 4526

Treffers et al (1994); imaged by HST

During

After

Subtracted

z=0.64

2003be


TYPING

Spectra: Type Ia have no H lines

Confirming Ia is easy with spectra.

Much harder with just photometry (i.e., low-res spectral information)

FOLLOW-UP

~ 15 days

~ 20 days

Flux

Guy et al (2007)

Lightcurves: time-evolution of brightness in several colour filters

MJD 53100 53150 53200

Flux

SNLS-04D3gx
Supernovae Type Ia: Origin

**PROGENITORS**

CO white dwarf accreting mass.

- **Single-degenerate** (e.g. Hosseinzaden et al, 2017)
- **Double-degenerate** (e.g. Roche & Garnavich, 2020)

**EXPLOSION**

C detonation creates thermonuclear explosion.

- Kruger et al (2012)

**LUMINOSITY**

- Lightcurve powered by radioactive decay of $^{56}\text{Ni}$.
- Higher core density $\Rightarrow$
- Larger mass of $^{56}\text{Ni}$ & IGEs $\Rightarrow$
- Higher luminosity & opacity $\Rightarrow$
- SN Ia brighter, slower to fade


Scatter: ~ 0.1 mag

Linear anti-correlation $\sim$ 0.3-0.7 mag

Peak Absolute Magnitude

- Brighter
- Fainter

Luminosity drop 15 days after peak

Slow $\leftrightarrow$ Fast
Standardization of SNIas

"Brighter SNIa are slow decliners"

Use the empirical 2D linear correlations between absolute magnitude and “stretch” and colour to standardise SNIas to within ~0.1 mag residual dispersion at peak

Before correction

After correction

Kim et al (2007)
Constraining Cosmological Parameters

DISTANCE-REDSHIFT RELATION

Measure redshift and distance modulus

\[ \mu = m_B - M + \alpha x_1 - \beta c \]

\[ \mu = 5 \log_{10} \frac{D_L}{\text{Mpc}} + 25 \]

DIFERENTIAL DISTANCE MEASUREMENT

Contours of constant \( D_L \) at various redshifts
COLLECT THE NOBEL PRIZE


The Nobel Prize in Physics 2011

Bayesian Hierarchical Modelling

**Prior**
- Standardization parameters
- Cosmological parameters
- Population parameters

**Intrinsic variability of SNIas**
- "True" values of observables
- Noise, selection effects
- Observed values

**SNIIa population distributions**
- Environmental properties (age, metallicity, SFR, host type)

Latent variables:
- unknown and unobserved, are integrated out during the inference

Noisy data subject to truncation

For more details, see:
- **BHM**: March, RT et al (2011)
- **BAHAMAS**: Shariff, RT et al (2016)
- **Simple-BayeSN**: Mandel et al (2017)
- **Steve**: Hinton et al (2019)
Supernovae Discoveries Over Time

- **1990s**: CCD cameras and robotic methods
- **1996**: Discovery of cosmic acceleration
- **O(10^3)** SNIa (spectroscopically confirmed)

Rubin Observatory - 10^6 SN/year
Statistical Challenge

To classify Type Ia vs non-Ia **reliably** and **efficiently** from light-curve data alone

**Classification challenges:**

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

Given a feature space, $X$, and a label space, $Y (K > 1 \text{ classes/dependent variables})$ we have $n_s$ labelled samples $\{x_i^s, y_i^s\}$ from the source domain $n_t$ unlabelled samples from the target domain, $\{x_i^t\}$.

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.
2015

Skyscrapers
Airplanes
Cars

Bikes
Gorillas
Graduation

Jacky lives on @jalce@playvicious.social now.
@jackyalce

2019

Cosmology

Incorrect classification of Type Ia vs non-Ia would lead to cosmological parameters systematic bias.

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.
Target Risk Estimation

Weighted ML estimation of risk

Approach: choose the classification/regression function $f(x)$ so as to minimise the risk (= expected loss) over the target domain.

Shimoidara (2000) showed:

$$E_{(x,y) \sim D}[\ell(f(x), y)] = E_{(x,y) \sim D_s} \left[ \frac{p_t(x)}{p_s(x)} \ell(f(x), y) \right]$$

The ratio of densities (weights) can be difficult to estimate reliably.

Bias correction approach

Let $s$ be a binary indicator variable controlling training set selection ($s=1$).

Zadrozny (2004) showed:

$$E_{(x,y) \sim D}[\ell(f(x), y)] = E_{(x,y) \sim \tilde{D}} \left[ \ell(f(x), y | s = 1) \right]$$

$$\tilde{D} = \frac{P(s = 1)}{P(s = 1 | x)} D$$

Estimate $p(s = 1 | x)$ via e.g. logistic regression, then draw samples from $\tilde{D}$.
Our Approach: Propensity Score Stratification

Work by Max Autenrieth (Stats PhD student), in collaboration with David van Dyk (Imperial)

Previous work (STACCATO) by Esben Revsbech (Revsbech, RT, van Dyk, 2018)

Propensity scores

\[ e(x_i) = \text{probability for object } i \text{ to be selected into the source domain:} \]

\[ e(x_i) \equiv P(s_i = 1 \mid x_s, x_t) \]

Idea:

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

\[ p_{s_j}(x, y) \approx p_{t_j}(x, y) \text{ for } j = 1, \ldots, k \]
Previous work: **SynThetically Augmented Light Curve ClassificATiOn (STACCATO)**, Revsbech et al (2018)

**Propensity score partitioning of target domain (test data)**

**Propensity score partitioning of source domain (training data)**

**Issue:** sparsity in lowest propensity score grouping requires augmentation via simulation in the source domain.

**Solution:** optimize GP-based augmentation via validation set (usually not available).
1. Fit Gaussian Process (GP) to light curve data
2. Apply diffusion map technique to map the fitted light curve into a covariate space of dim~O(100), following Richards et al (2012)
3. Stratify light curve data according to propensity scores quantiles
4. Sample new synthetic light curves from fitted GPs in groups with sparse training data (requires validation set to optimise augmentation scheme)
5. Use Random Forest on the stratified diffusion coordinates (group-by-group) for classification
Classification Performance with Augmentation

Representative source data (unbiased)

\[
\text{AUC} = 0.977
\]

STACCATO on biased source data

\[
\text{AUC} = 0.961
\]

Dashed (solid) lines are before (after) augmentation; colours are for individual strata; black combined.

STACCATO performance close to “gold standard” of unbiased training set (left)

Later, the STACCATO framework was improved and extended (w/o using validation) in AVOCADO (Boone, 2019), winner of the PLASTiCC challenge 2019.
Current Work: Stratified Learning

We seek a principled framework for covariate shift adaptation via propensity score stratification that does not need augmentation:
1. Augmentation is problem-specific
2. Augmentation usually requires a validation set (not available)

Propensity score stratification leads to approximately balanced (i.e., unbiased) sub-groups, on which to perform supervised classification/regression.

Work by Max Autenrieth (Stats PhD student), in collaboration with David van Dyk (Imperial)

Classification (2-way)
SNIa vs non-Ia, AUC:

Gold standard: 0.977
Best-in-class : 0.937
STACCATO : 0.961
StratLearning : **0.973**

Multivariate regression
Photo-z estimation
(~500,000 source/targets from SDSS DR 8)
Stratified Covariate Balancing

The improvement in covariate balancing between source and target domain is assessed via absolute standardised mean differences (smd) in each “stratum” $j$ and covariate $x_j$: 

\[ \text{smd}_j \equiv \frac{|\bar{x}_{s,j} - \bar{x}_{t,j}|}{\sqrt{\frac{\sigma^2_{s,j} + \sigma^2_{t,j}}{2}}} \]

$\bar{x}_{\tau,j}$ = sample mean for $\tau = s, t$

$\sigma_{\tau,j}$ = sample std for $\tau = s, t$

Balance improves within stratum
Conditional Density Estimation Problem

Work by Max Autenrieth (Stats PhD student), in collaboration with David van Dyk (Imperial)

The problem: estimate the conditional density of redshift, $z$, given the observed covariates (magnitudes in 5 different colour filters), in the presence of covariate shift.
We compare our performance to the estimators in Izbicki et al (2017).

Approach:

1. “StratLearning” partitions the source and target domain according to propensity scores (no augmentation needed)
2. Within each group, we combine two conditional density estimation models from Izbicki et al (2017), ker-NN and Series, via a weighted average.
3. Weight is optimised by minimising the empirical loss on a validation set (a sub-set of the training set, no test data needed)
“StratLearning”: Photo-z Performance

Moderate shift
Low covariate dimensions

Moderate shift
High covariate dimensions

StratLearning outperforms previous methods for this problem. Performance improvement is larger in the presence of high-D noisy covariate space (right panel).

Work by Max Autenrieth (Stats PhD student), in collaboration with David van Dyk (Imperial). Paper upcoming.
Conclusions

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 based on stratifying source and target domain according to propensity scores (= probability of an object to be included in the source domain). Methods paper upcoming.

3. Within strata, source and target domains are better balanced: “Stratified Learning” 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.)
Thank you!

www.robertotrotta.com
@R_Trotta

Currently on leave of absence in Trieste, Italy

Starting up a new Data Science Initiative
References (Astro)


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.


