

Supernova classification by decision tree ensembles and automated machine learning

The Dark Energy Revolution in Cosmology
Celebrating Ioav Waga's Career

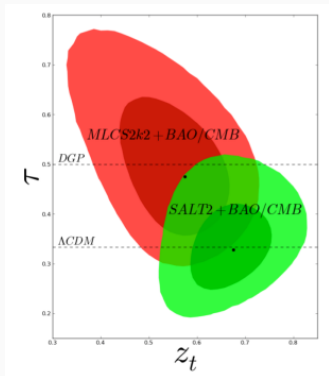
Marcelo Vargas dos Santos, Ribamar Reis, Miguel Quartin

Rio de Janeiro, 28 de Agosto de 2018

Universidade Federal do Rio de Janeiro

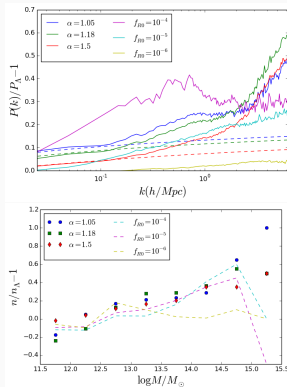
Why am I here?

Master (2010-2012):
 $q(z)$ Cosmology



Tension: MLCS2k2 and SALT2
Giostri et. al. 2012

PhD (2012-2016):
Modified Gravity



Numerical Simulations for γ -gravity
Santos, M. V. et. al. 2016

The best advice

*Some advisors train their students to be copies of themselves.
I do not think like that, you must develop your own skills.
- Ioan Waga*

The best advice

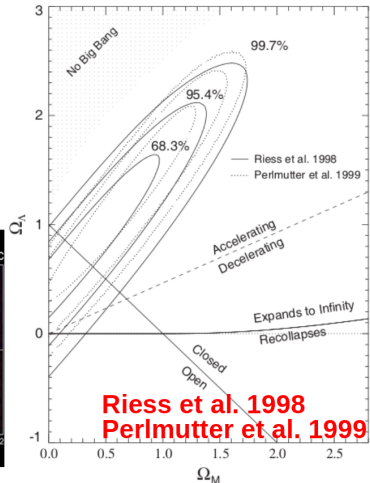
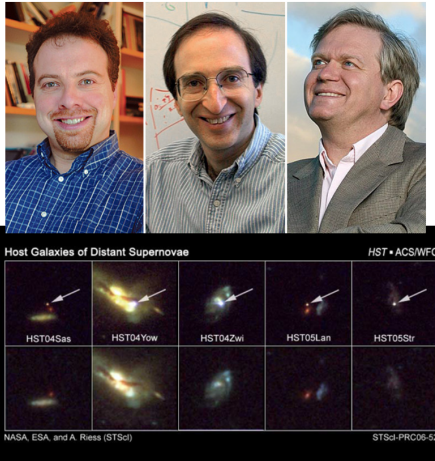
*Some advisors train their students to be copies of themselves.
I do not think like that, you must develop your own skills.*

- Ioan Waga

So, Mathematica is only for plots. Let me work with C++.

- Me

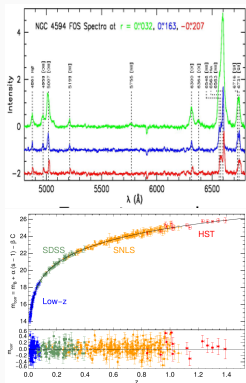
Type Ia Supernovae



Distances by Type Ia Supernovae (No H lines + Sill line)

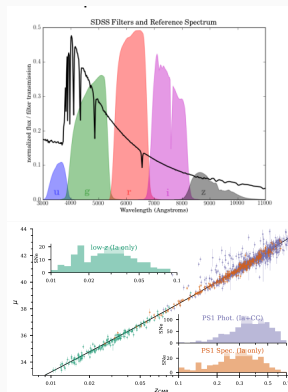
Spectroscopy vs. Photometry

Spectroscopy



~ 700 JLA (Confirmed by spectra)
Betoule et. al. 2014

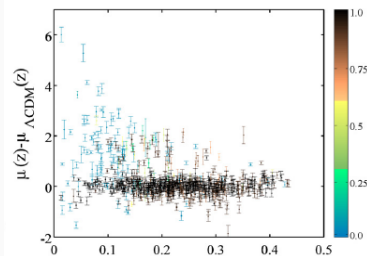
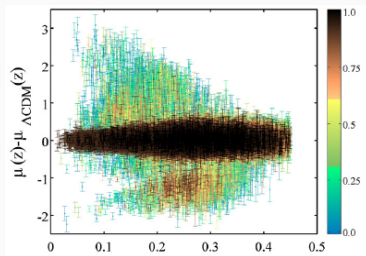
Photometry



Pantheon
(~ 1000 spec + ~ 200phot)
Scolnic et. al. 2017

Why is a good classification important?

A contaminated set may introduce errors and bias on the measures.



Hlozek et al. 20124

Examples:

- Mag increases $\Rightarrow \Omega_{m0}$ (ΛCDM) decreases
- Distribution on redshift $\Rightarrow w_{DE} \neq -1$ or $w_{DE} = w(z)$
- Non-gaussian moments distribution $\Rightarrow$ Fake lensing

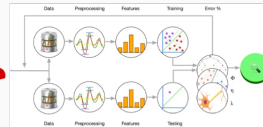
Machine Learning

Training set

id	x_1	x_2	...	x_n	y_{obs}
1	x_{11}	x_{21}	...	x_{n1}	y_1
2	x_{12}	x_{22}	...	x_{n2}	y_2
3	x_{13}	x_{23}	...	x_{n3}	y_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_{1N}	x_{2N}	...	x_{nN}	y_N

x_i : features (photometry)

y_{obs} : target (spectroscopy)



Prediction set

id	x_1	x_2	...	x_n
1	x_{11}	x_{21}	...	x_{n1}
2	x_{12}	x_{22}	...	x_{n2}
3	x_{13}	x_{23}	...	x_{n3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_{1N}	x_{2N}	...	x_{nN}

x_i : features (photometry)

y_{pred}
y_1
y_2
y_3
$\vdots$
y_N

machine
learning

Features: Light curve

Target: Ia or Core Colapse (Binary Classification)

Supernova Photometric Classification Challenge

The challenge: Kessler et. al. 2010:

- 21319 Simulated Light curves
- DES Filters (*griz*)
- Types: Ia: 4459, Ibc: 2299, II: 11384
- Host galaxy redshift provided (z_{host})
- Training set ~ 1100
- Results: ArXiv:1008.1024

Supernova Photometric Classification Challenge

Revisited: Lochner et. al. 2016

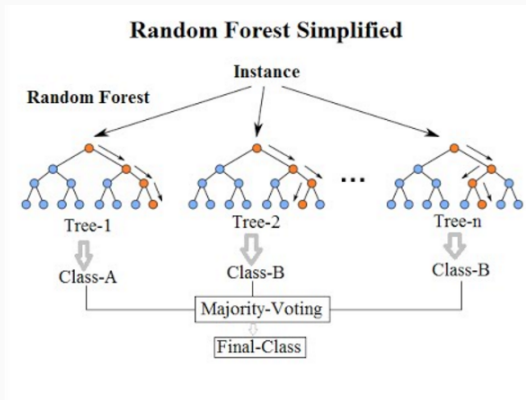
- 4 Features Sets: Salt2, Newling, Karpenka, Wavelets
- Classification by Machine Learning: Naive Bayes, K Nearest Neighbors, Artificial Neural Networks, Support Vector Machines, Boosted Decision Trees
- Metric: Area under the ROC curve
- Results: BDT (ADA) best method; SALT2 > Wavelets > Newling > Karpenka

Supernova Photometric Classification Challenge

Revisiting again: Our work

- 4 Features Sets: Salt2, Newling, Karpenka, Wavelets
- Classification by Machine Learning: Decision trees essembles and Automated Machine Learning
- Metric: Area under the ROC curve and the Average Precision
- Train for 200 different training sets (Randomized search CV)

Decision Trees Ensembles

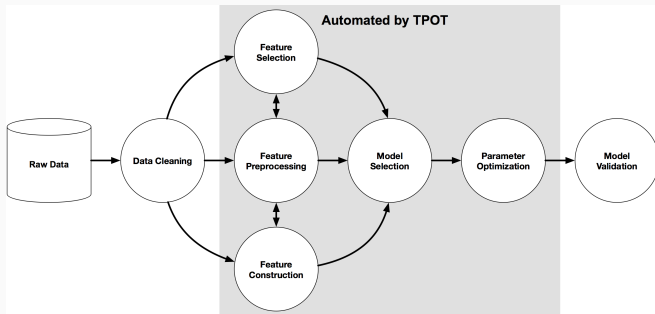


Average methods: Extra Trees (**EXT**) and Random Forest (**RF**)

Boosting methods: Ada Boost (**ADA**), Gradient Boost (**GB**), Extreme Gradient Boost (**XGB**)

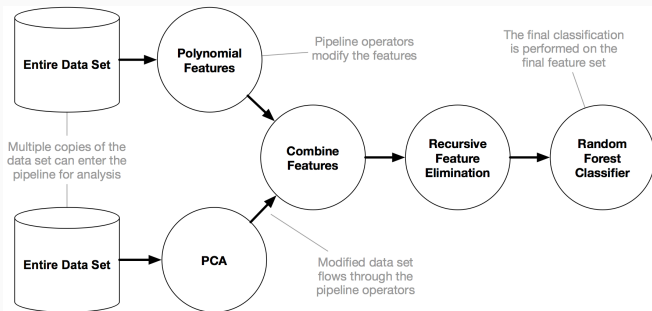
Python libraries: Scikit-Learn and XGBoost

Automated Machine Learning (TPOT)



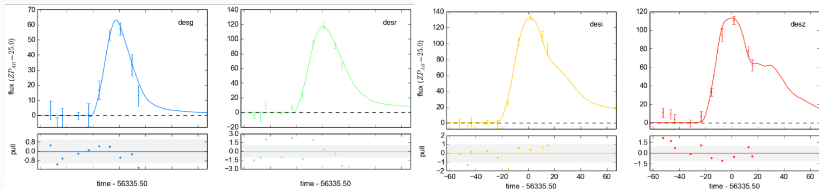
Pipeline determined by genetic optimization

Automated Machine Learning (TPOT)



Library: <http://epistasislab.github.io/tpot/>

Template fitting (SALT2)



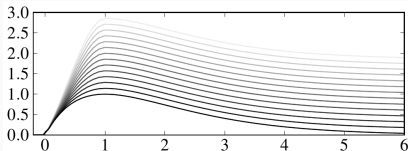
$$F(t, \lambda) = x_0 \times [M_0(t - t_0, \lambda) + x_1 M_1(t - t_0, \lambda)] \times \exp[c \times CL(\lambda)],$$

Features: t_0 , x_0 , x_1 , c and z (Fitting with PyMultinest).

Fitting: PyMultinest

Parametric Models (Newling)

Newling



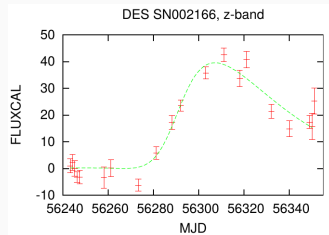
$$F(t) = A \left(\frac{t - \phi}{\sigma} \right)^k \exp \left(- \frac{t - \phi}{\sigma} \right) k^{-k} e^k + \Psi(t)$$

Features per filter:

A, ϕ, σ, k, Ψ

Total: 20 features.

Karpenka



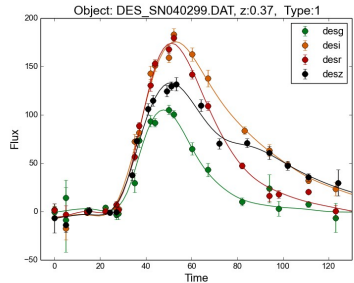
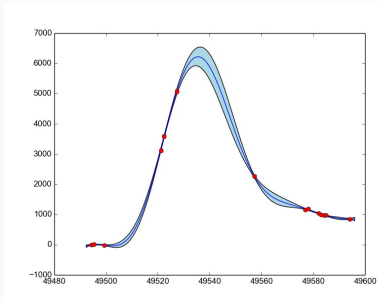
$$F(t) = A[1 + B(t - t_1)^2] \frac{\exp[-(t - t_0)/T_{\text{fall}}]}{1 + \exp[-(t - t_0)/T_{\text{rise}}]}$$

Features per filter:

$t_0, t_1, T_{\text{fall}}, T_{\text{rise}}, A, B.$

Total: 24 features.

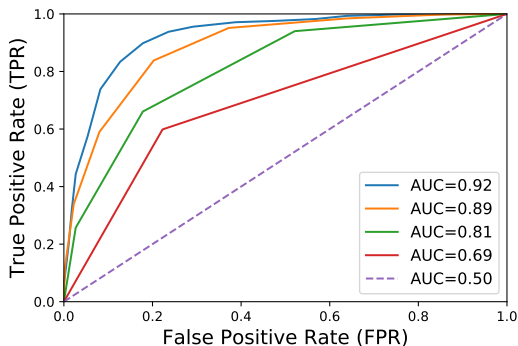
Wavelets decomposition



Gaussian Process Fit $\Rightarrow$ Wavelet decomposition $\Rightarrow$ PCA

Features: 20 Principal components

Metrics - Receiver Operating Characteristic (ROC) Curve



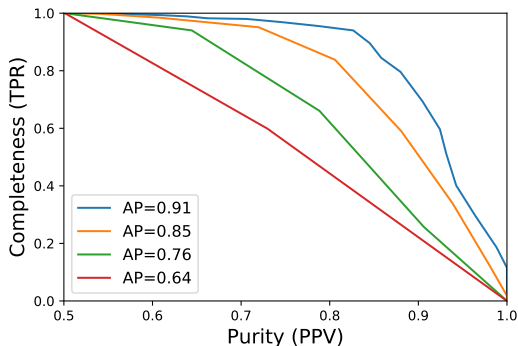
$$TPR = \frac{tp}{tp + fn}$$

$$FPR = \frac{fp}{fp + tn}$$

AUC = Area Under the Curve

Perfect situation: $TPR = 1$ and for any $FPR > 0 \Rightarrow AUC = 1$

Metrics - Completeness vs Purity (Precision-Recall)



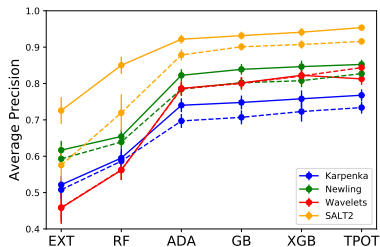
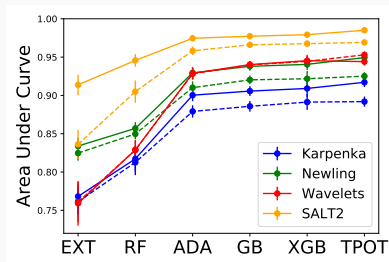
$$\text{TPR} = \frac{tp}{tp + fn}$$

$$\text{PPV} = \frac{tp}{tp + fp}$$

$$AP = \text{Average Precision} = \sum_i (\text{TPR}_{i+1} - \text{TPR}_i) \text{PPV}_i$$

Perfect situation: $\text{PPV} = 1$ for any $\text{TPR} \Rightarrow AP = 1$

Results - Scores



High scores:

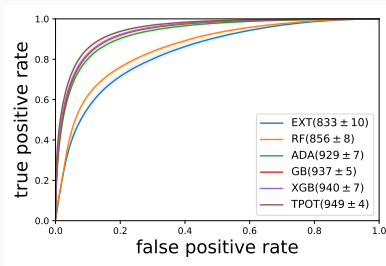
Features: SALT2, but maybe biased; ML: Boosting models + **TPOT**

Low scores:

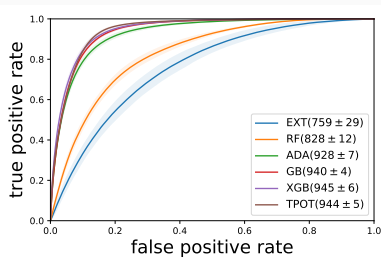
Karpenka; ML; Averaging models

Results - ROC Curve

Newling

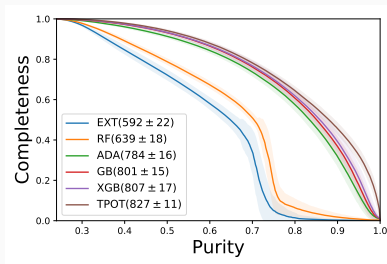


Wavelets

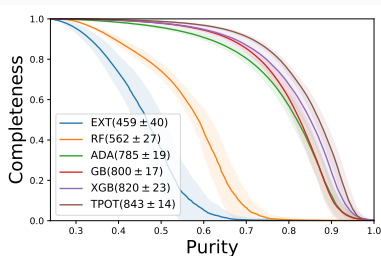


Results - Completeness-Purity

Newling



Wavelets



- The best scores comes from the boosted ensembles: **ADA**, **GB** and **XGB** in any situation.
- The scores from **TPOT** are competitive with the boosted models.
- In the comparison between the features sets the SALT2 reach the best scores. However many supernovae from the original data had been simulated by this model, so this result is possibly biased.
- In the comparison between Wavelets and Newling, the second one is better, because it can produce larger pure sets.
- Paper (**soon**): The results for the training sets of 550 and 5500 data, impact over the residual plot, and many many other plots ...



Thank you, Ioav.