Machine Learning in Astronomy

Reza Monadi

UC Riverside

May 14, 2020

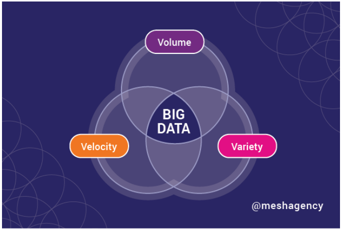
- Astronomy in **Big Data** era

- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy

- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy
- Implementation of unsupervised **ML** in astronomy

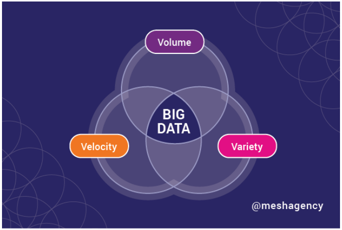
- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy
- Implementation of unsupervised **ML** in astronomy
- Limitations of **ML** in astronomy

3Vs in astronomy



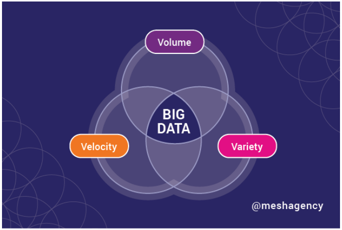
- Volume: larger quantities of data by the advent of better telescopes and surveys

3Vs in astronomy



- Volume: larger quantities of data by the advent of better telescopes and surveys
- Velocity: Higher speed of incoming observational data

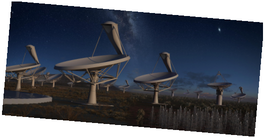
3Vs in astronomy

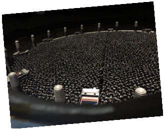


- Volume: larger quantities of data by the advent of better telescopes and surveys
- Velocity: Higher speed of incoming observational data
- Variety: multi-wavelength spectroscopic and photometeric data from versatile astronomical objects

Astronomical surveys are Astronomicaly growing

Sky Server Prokects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	40 PB expected
LSST (The Large Synoptic Survey Telescope)	200 PB expected
SKA (The Square Kilometer Array)	4.6 EB expected

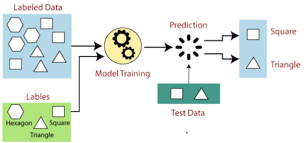
SKA


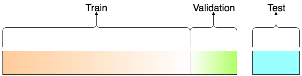
DESI


ZFT


LSST

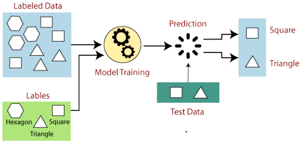

Stages of Supervised Learning

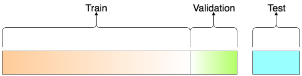




- Training:

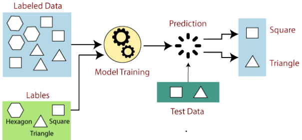
Stages of Supervised Learning

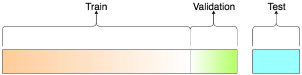




- Training:
  1. Select a model

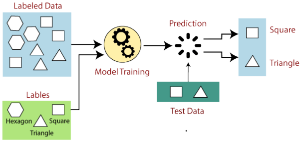
Stages of Supervised Learning

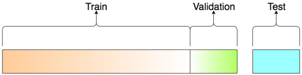




- Training:
  1. Select a model
  2. Set up hyper-parameters of model

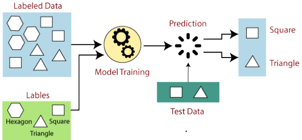
Stages of Supervised Learning

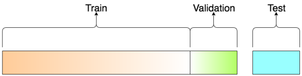




- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set

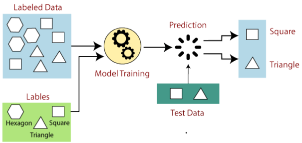
Stages of Supervised Learning

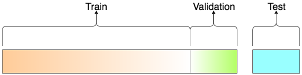




- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set

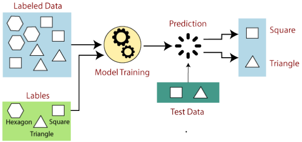
Stages of Supervised Learning

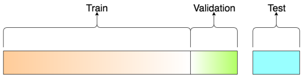




- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
  4. Validate the hyper-parameters

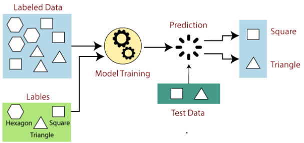
Stages of Supervised Learning

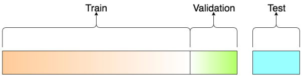




- Training:
    1. Select a model
    2. Set up hyper-parameters of model
    3. Teach the machine by training set
    4. Validate the hyper-parameters
    5. Select the optimum hyper-parameters

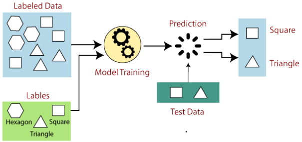
Stages of Supervised Learning

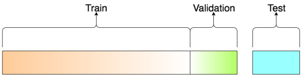




- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
  4. Validate the hyper-parameters
  5. Select the optimum hyper-parameters
- Testing:

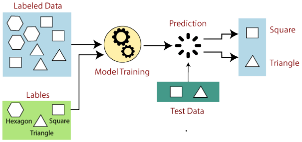
Stages of Supervised Learning

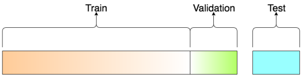




- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
  4. Validate the hyper-parameters
  5. Select the optimum hyper-parameters
- Testing:
  1. Test learned model by an unseen part of the data-set.

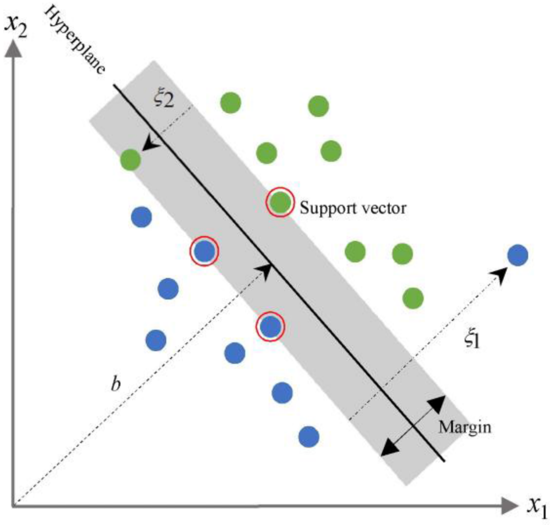
Stages of Supervised Learning





- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
  4. Validate the hyper-parameters
  5. Select the optimum hyper-parameters
- Testing:
  1. Test learned model by an unseen part of the data-set.
  2. Select the best model and use it for predictions.

How Support Vector Machine Works?



Classifying Pre-Main-Sequence Stars using SVM

Hubble Tarantula Treasury Project – VI. Identification of Pre–Main-Sequence Stars using Machine Learning techniques

Victor F. Ksoll,[1,2,⋆] Dimitrios A. Gouliermis,[1,3,†] Ralf S. Klessen,[1] Eva K. Grebel,[4] Elena Sabbi,[5] Jay Anderson,[5] Daniel J. Lennon,[6] Michele Cignoni,[7] Guido de Marchi,[8] Linda J. Smith,[9] Monica Tosi,[10] and Roeland P. van der Marel[5]

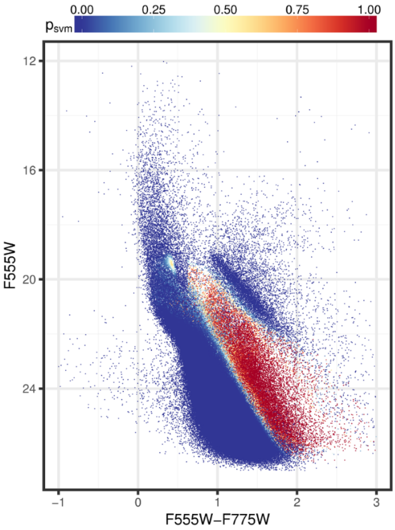
[1] Institut für Theoretische Astrophysik, Zentrum für Astronomie der Universität Heidelberg, Albert-Ueberle-Str. 2, 69120 Heidelberg, Germany
[2] Interdisciplinary Center for Scientific Computing, University of Heidelberg, Mathematikon, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[3] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[4] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, 69120 Heidelberg, Germany
[5] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[6] ESA – European Space Astronomy Center, Apdo. de Correo 78, E-28691 Associate Villanueva de la Cañada, Madrid, Spain
[7] Department of Physics, University of Pisa, Largo Pontecorvo 3, I-56127 Pisa, Italy
[8] European Space Research and Technology Centre, Keplerlaan 1, 2200 AG Noordwijk, Netherlands
[9] European Space Agency and Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[10] INAF-Osservatorio Astronomico di Bologna, Via Ranzani 1, I-40127 Bologna, Italy

Draft version 21 May 2018

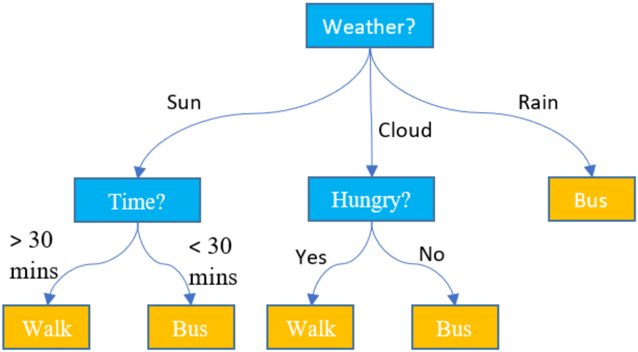
ABSTRACT

The Hubble Tarantula Treasury Project (HTTP) has provided an unprecedented photometric coverage of the entire star-burst region of 30 Doradus down to the half Solar mass limit. We use the deep stellar catalogue of HTTP to identify all the pre–main-sequence (PMS) stars of the region, i.e., stars that have not started their lives on the main-sequence yet. The photometric distinction of these stars from the more evolved populations is not a trivial task due to several factors that alter their colour-magnitude diagram positions. The identification of PMS stars requires, thus, sophisticated statistical methods. We employ Machine Learning Classification techniques on the HTTP survey of more than 800,000 sources to identify the PMS stellar content of the observed field. Our methodology consists of 1) carefully selecting the most probable low-mass PMS stellar population of the star-forming cluster NGC 2070, 2) using this sample to train classification algorithms to build a predictive model for PMS stars, and 3) applying this model in order to identify the most probable PMS content across the entire Tarantula Nebula. We employ Decision Tree, Random Forest and Support Vector Machine classifiers to categorise the stars as PMS and Non-PMS. The Random Forest and Support Vector Machine provided the most accurate models, predicting about 20,000 sources with a candidateship probability higher than 50 percent, and almost 10,000 PMS candidates with a probability higher than 95 percent. This is the richest and most accurate photometric catalogue of extragalactic PMS candidates across the extent of a whole star-forming complex.

Classifying Pre-Main-Sequence Stars using SVM



How Decision Tree works?



Using DT for classifying galaxies and stars in SDSS?

THE ASTRONOMICAL JOURNAL, 141:189 (12pp), 2011 June doi:10.1088/0004-6256/141/6/189

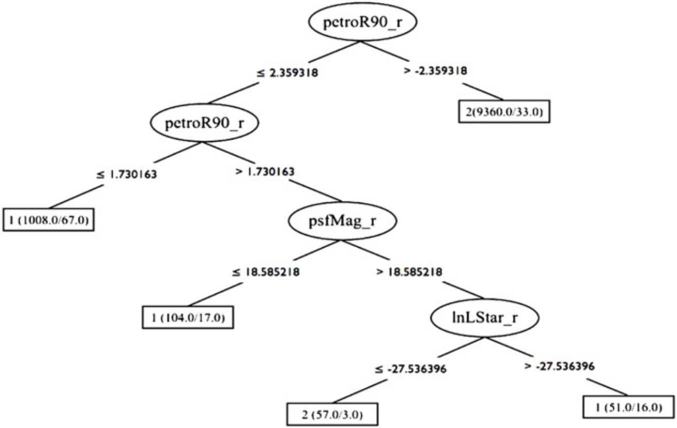
DECISION TREE CLASSIFIERS FOR STAR/GALAXY SEPARATION

E. C. VASCONCELLOS[1], R. R. DE CARVALHO[2], R. R. GAL[3], F. L. LABARBERA[4],
H. V. CAPELATO[2], H. FRAGO CAMPOS VELHO[5], M. TREVISAN[6], AND R. S. R. RUIZ[1]
[1] CAP, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[2] DAS, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[3] Institute for Astronomy, University of Hawaii, 2680 Woodlawn Dr., Honolulu, HI 96822, USA
[4] INAF-Osservatorio Astronomico di Capodimonte, via Moiariello 16, Napoli 80131, Italy
[5] LAC, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[6] IAG, University of São Paulo, Rua do Matão 1226, São Paulo 05508-090, Brazil
Received 2010 October 27; accepted 2011 February 11; published 2011 May 9

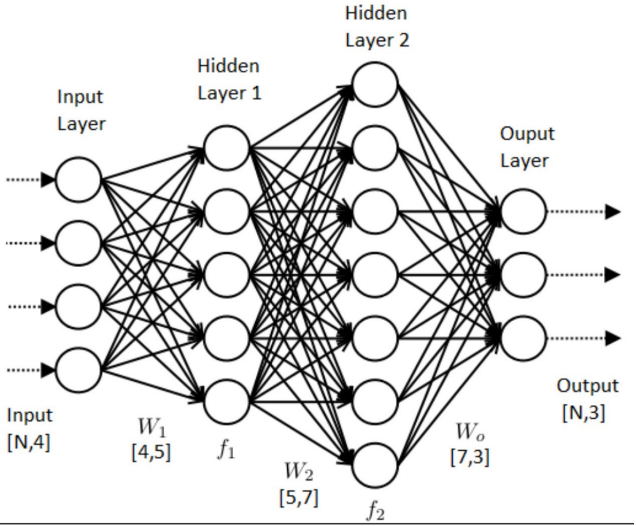
ABSTRACT

We study the star/galaxy classification efficiency of 13 different decision tree algorithms applied to photometric objects in the Sloan Digital Sky Survey Data Release Seven (SDSS-DR7). Each algorithm is defined by a set of parameters which, when varied, produce different final classification trees. We extensively explore the parameter space of each algorithm, using the set of 884,126 SDSS objects with spectroscopic data as the training set. The efficiency of star–galaxy separation is measured using the completeness function. We find that the Functional Tree algorithm (FT) yields the best results as measured by the mean completeness in two magnitude intervals: $14 \leqslant r \leqslant 21$ (85.2%) and $r \geqslant 19$ (82.1%). We compare the performance of the tree generated with the optimal FT configuration to the classifications provided by the SDSS parametric classifier, 2DPHOT, and Ball et al. We find that our FT classifier is comparable to or better in completeness over the full magnitude range $15 \leqslant r \leqslant 21$, with much lower contamination than all but the Ball et al. classifier. At the faintest magnitudes ($r > 19$), our classifier is the only one that maintains high completeness (>80%) while simultaneously achieving low contamination (~2.5%). We also examine the SDSS parametric classifier (psfMag – modelMag) to see if the dividing line between stars and galaxies can be adjusted to improve the classifier. We find that currently stars in close pairs are often misclassified as galaxies, and suggest a new cut to improve the classifier. Finally, we apply our FT classifier to separate stars from galaxies in the full set of 69,545,326 SDSS photometric objects in the magnitude range $14 \leqslant r \leqslant 21$.

Using DT for classifying galaxies and stars in SDSS?



How Neural Network works?



Convolutional Neural Network and DLAs

Deep Learning of Quasar Spectra to Discover and Characterize Damped Lyα Systems

David Parks,[1] J. Xavier Prochaska,[2] Shawfeng Dong,[3] Zheng Cai,[2,4]

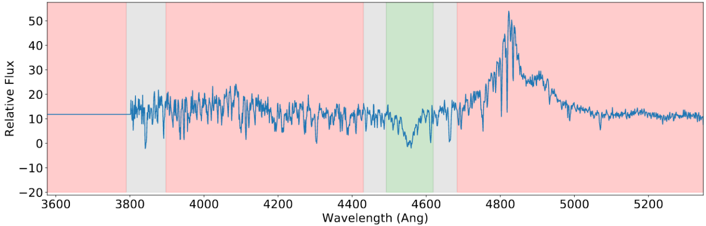
[1] *Computer Science, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA — dfparks@ucsc.edu*
[2] *Astronomy & Astrophysics, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA*
[3] *Applied Mathematics & Statistics, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA*
[4] *Hubble Fellow*

Accepted XXX. Received YYY; in original form ZZZ

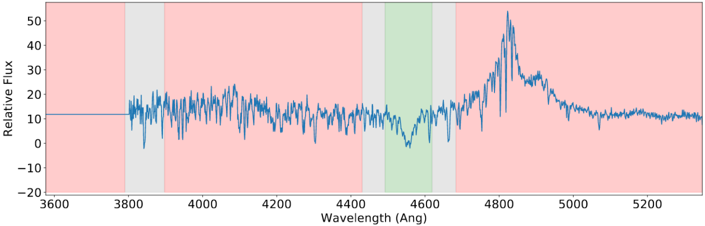
ABSTRACT

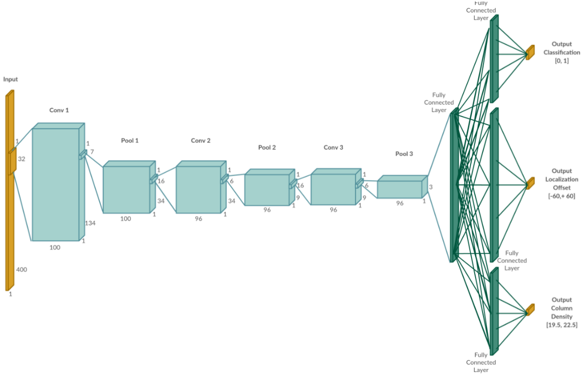
We have designed, developed, and applied a convolutional neural network (CNN) architecture using multi-task learning to search for and characterize strong HI Lyα absorption in quasar spectra. Without any explicit modeling of the quasar continuum nor application of the predicted line-profile for Lyα from quantum mechanics, our algorithm predicts the presence of strong HI absorption and estimates the corresponding redshift $z_{\rm abs}$ and HI column density $N_{\rm HI}$, with emphasis on damped Lyα systems (DLAs, absorbers with $N_{\rm HI} \geq 2 \times 10^{20}\,{\rm cm}^{-2}$). We tuned the CNN model using a custom training set of DLAs injected into DLA-free quasar spectra from the Sloan Digital Sky Survey (SDSS), data release 5 (DR5). Testing on a held-back validation set demonstrates a high incidence of DLAs recovered by the algorithm (97.4% as DLAs and 99% as an HI absorber with $N_{\rm HI} > 10^{19.5}\,{\rm cm}^{-2}$) and excellent estimates for $z_{\rm abs}$ and $N_{\rm HI}$. Similar results are obtained against a human-generated survey of the SDSS DR5 dataset. The algorithm yields a low incidence of false positives and negatives but is challenged by overlapping DLAs and/or very high $N_{\rm HI}$ systems. We have applied this CNN model to the quasar spectra of SDSS-DR7 and the Baryonic Oscillation Spectroscopic Survey (BOSS, data release 12) and provide catalogs of 4,913 and 50,969 DLAs respectively (including 1,659 and 9,230 high-confidence DLAs that were previously unpublished). This work validates the application of deep learning techniques to astronomical spectra for both classification and quantitative measurements.

Training DLA/noDLA models: 1D image recognition

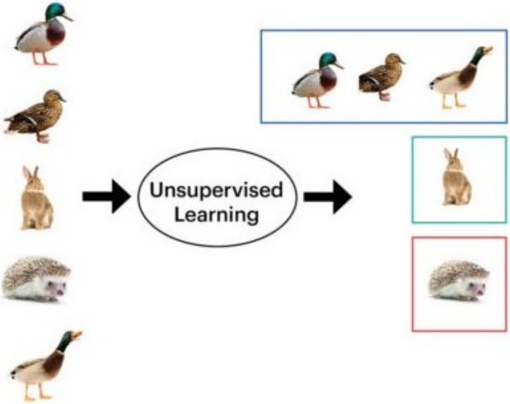


Training DLA/noDLA models: 1D image recognition





How unsupervised learning works?



Unsupervised Learning and knowledge discovery

- Clustering:
  Distinguishing similar objects and patterns in the data-set

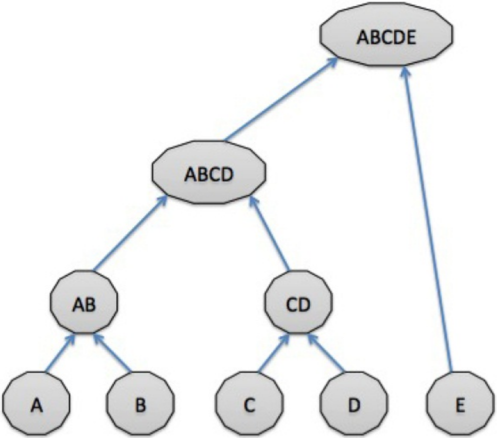
Unsupervised Learning and knowledge discovery

- Clustering:
  Distinguishing similar objects and patterns in the data-set
- Anomaly detection:
  Recognizing odd objects which cannot be categorized and do not belong to any group in the data-set

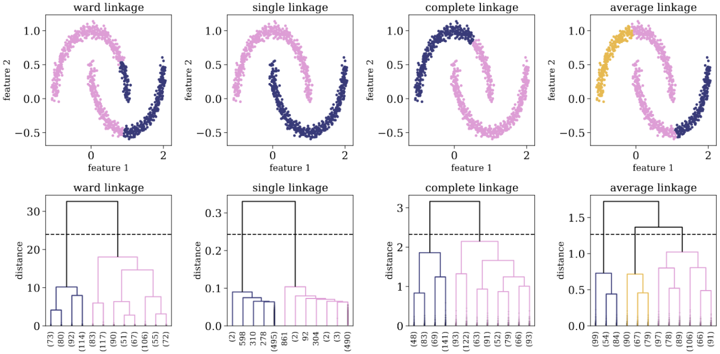
Unsupervised Learning and knowledge discovery

- Clustering:
  Distinguishing similar objects and patterns in the data-set
- Anomaly detection:
  Recognizing odd objects which cannot be categorized and do not belong to any group in the data-set
- Visualizationand dimentionality reduction:
  Maping high dimensional data set to a 2D or 3D space

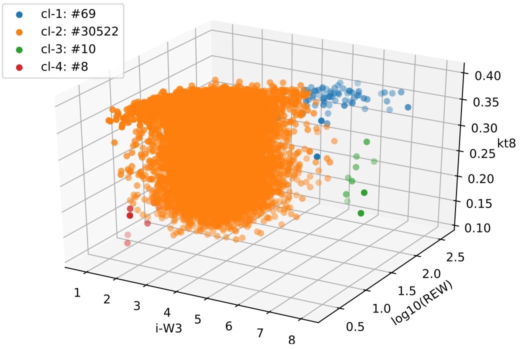
How Agglomerative clustering works?



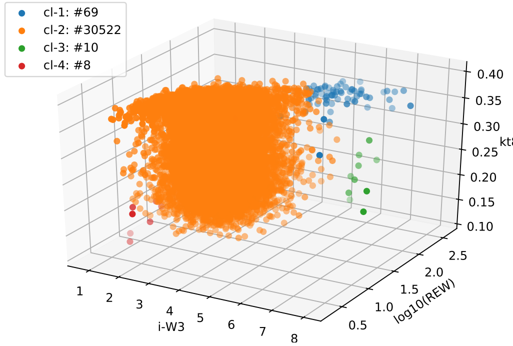
Linkage methods in Agglomerative clustering?

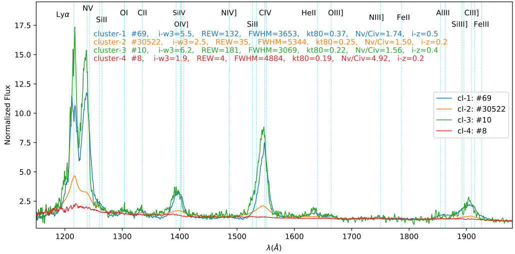


Agglomerative clustering finds weird quasars?

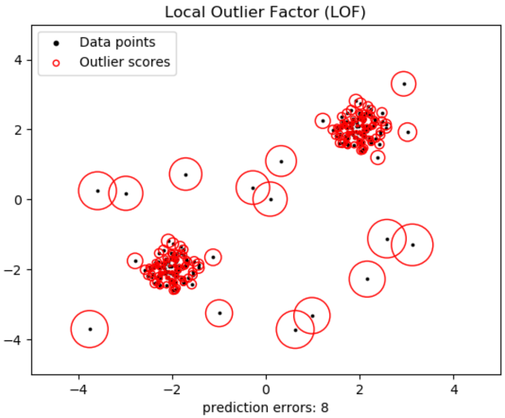


Agglomerative clustering finds weird quasars?

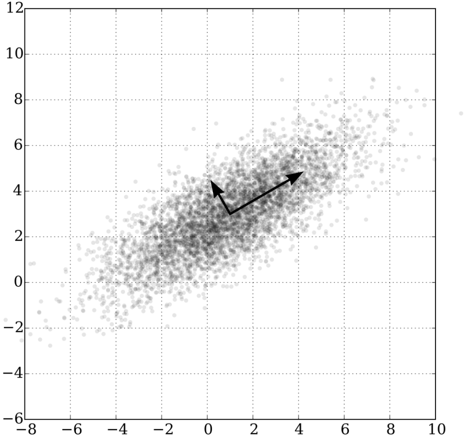




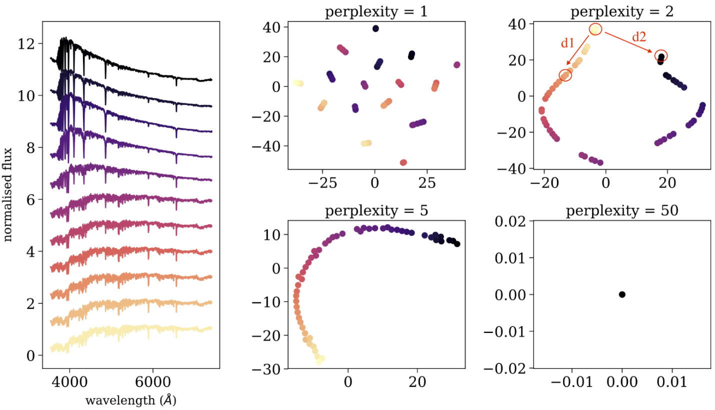
Local Outlier Factor finds outliers



Dimensionality Reduction with Principal Component Analysis



Visualization with t-Distributed Stochastic Neighbor Embedding



Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results

Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results
- ML algorithms are mostly designed for industry not science.

Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results
- ML algorithms are mostly designed for industry not science.
- Training on a biased sample is very dangerous

Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results
- ML algorithms are mostly designed for industry not science.
- Training on a biased sample is very dangerous
- ML is not plug and play

Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results
- ML algorithms are mostly designed for industry not science.
- Training on a biased sample is very dangerous
- ML is not plug and play
- There are a variety of methods for each task

Limitation of ML in astronomy

- Lack of Pre-procession and normalization can end up with misleading/wrong results
- ML algorithms are mostly designed for industry not science.
- Training on a biased sample is very dangerous
- ML is not plug and play
- There are a variety of methods for each task
- Some algorithms scarify interpretablity for functionality

References

1. **ML, General:** Baron, Dalya. "Machine learning in astronomy: A practical overview." arXiv preprint arXiv:1904.07248 (2019).
2. **Big Data:** Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the big data era." Data Science Journal 14 (2015).
3. **SVM:** Ksoll, Victor F., et al. "Hubble Tarantula Treasury ProjectVI. Identification of pre-main-sequence stars using machine-learning techniques." Monthly Notices of the Royal Astronomical Society 479.2 (2018): 2389-2414.
4. **Decision Tree:** Vasconcellos, E. C., et al. "Decision tree classifiers for star/galaxy separation." The Astronomical Journal 141.6 (2011): 189.
5. **Neural Network:** Parks, David, et al. "Deep learning of quasar spectra to discover and characterize damped Ly systems." Monthly Notices of the Royal Astronomical Society 476.1 (2018): 1151-1168.
6. **Agglomerative clustering:** R. Monadi, F. Hamann, S. Bird, Precise Selection of Extremely Red Quasars, (in preparation)


THANK YOU
FOR YOUR KIND ATTENTION