data science for (physical) scientists I

I: what is data science and working environment

dr.federica bianco | *fbb.space* | *fedhere* | *fedhere*

1 what is data science

2 the scientific method

epistemology

falsifiability

probabilistic induction

reproducibility

3 data science tools

github

python

jupyter notebooks

google colab

stackoverflow

this slide deck

slides.com/federicabianco/dsps_01

data science for (physical) scientists I

dr.federica bianco | *fbb.space* | *fedhere* | *fedhere*

1

some administrative stuff

Syllabus

https://bit.ly/dspsSyllabus

Learning Outcomes

By the end of this class you should be able to formulate an appropriate analysis plan for a research question, select, gather, and prepare data for analysis, and choose and apply machine learning methods to the data.

Syllabus

https://bit.ly/dspsSyllabus

Data Science for Physical Scientists Fall 2023 - PHYS4/661

Instructor: Dr. Federica B. Bianco (she/her) fbianco@udel.edu
TAs: Willow Fox Fortino (she/her) fortino@udel.edu

Code of Conduct: Diversity is considered a resource that enriches us culturally and intellectually in this class. No instances of harassment or attempts to marginalize students will be tolerated in my class. Be respectful and collaborate instead of competing. If you have concerns please come talk to me.

READ THE FULL CODE OF CONDUCT https://bit.ly/dspsCoC

UD resources and Professional Conduct Statement:

UD Diversity Equity and Inclusion resources

The office for Equity and Inclusion (OEI) at UD https://sites.udel.edu/oei/ is the appropriate place to report when the report is intended to trigger an investigation. The office also provides training and consultation for DEI-related matters.
The UD Office of Diversity and Inclusion has developed a strategic plan https://sites.udel.edu/diversity/
DPA Climate, Diversity, Equity, and Inclusion Committee: http://dpacdeicommittee.github.io/

Counterspaces accessible at UD:

The UD Diversity Caucuses https://sites.udel.edu/oei/diversity-caucuses/
UD Black Students Union http://udel.edu/stu-org/bsu/bsuhome.htm
UD BLACK GRADUATE STUDENT ASSOCIATION https://sites.udel.edu/bgsa/
UD Pride Caucus https://sites.udel.edu/pride

Syllabus

https://bit.ly/dspsSyllabus

- 5% pre-class questions
- 15% class participation (get up and code!!)
- 25% homework
- 15% midterm
- 30% final

Syllabus

https://bit.ly/dspsSyllabus

- 5% pre-class questions
- 15% class participation (get up and code!!)
- 25% homework
- 15% midterm
- 30% final

from beginning of class to 5 minutes past (be on time!)

questions on previous class material and reading assignments

Syllabus

https://bit.ly/dspsSyllabus

- 5% pre-class questions
- 15% class participation (get up and code!!)
- 25% homework
- 15% midterm
- 30% final

ask questions

answer questions

get up and code

extra credit assignments

Syllabus

https://bit.ly/dspsSyllabus

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

Homework projects must be turned in as *jupyter notebooks* by checking them into your github account in a DSPS_<firstinitialLastname> repo and the project directories HW<hw number> (unless otherwise stated).

<finitialLastname> is e.g. fBianco

grading the homework

homework are assigned as skeleton notebooks with missing code

you will be graded on

(80% of the grade)

1) rendered plots and their captions

2) obtaining correct numbers

3) interpreting each result you get

- **5% pre-class questions**
- **15% class participation**
- **25% homework**
- **15% midterm**
- **30% final**

The target variables for our analysis are redshift and brightness: "zcmb" and "mb". This an exercise about fitting lines to data. Why does it lot look like there is a line relationship between them? How can does the relationship look? Plot the "zcmb" vs "mb". Include the y uncertainty which is reported in "dmb". Plot each data point with a different color based on the survey it comes from. The survey data is indicated by the variable "set". To do it you can use the dataframe plotting methods (`df.plot.scatter()`), or pylab: `pylab.scatter()`, or `pylab.plot()` --asking to plot the data as point ('.')-- but for each the way you relate the color to a column value is a bit different. Stackoverflow is your friend here!

In [0]:

It is a **power law**, which means it will be linear in logarithmic space - specif

Replot the data with natural y axis and logarithmic x axis, you can do th `df.plot.scatter()` you can use `logx=True`)

Note that when we talk about log we (generally) talk about log base 10!

```
#scatter plot of the z-shift zcmb and brightness mb of supernovaes
ax = jla.plot.scatter("zcmb","mb",yerr="dmb",s=40,c= "set",colormap='vi
ridis', figsize=(15,15))
ax.set(xlabel='zcmb', ylabel='mb');
```

Figure 2: Supernovae brightness plotted against redshift z. Errorbars are shown for each data points. Different color of points indicate different sets of data. The plot shows power law between brightness and redshift of supernovae.

homework

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

Please work in groups of up to 3-5 people on homework as a collaborative projects.

Each group will turn in one homework, but all members of the group are responsible for the assignment.

The assignment must be uploaded in every student's repository. Does not have to be identical for all group members, but it has to be just as complete as the one turned in

homework

A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example these contributions).

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

nature

Contributions

All authors contributed to the drafting of the paper. A.R., N.S. and R.C.S. imaged the area around η Car. A.R. and M.E.H. reduced the imaging data. H.E.B. provided images of the echoes that guided our spectroscopic pointings. J.L.P., R.C., R.J.F. and W.F. obtained the spectra and reduced them. A.R. and J.L.P. performed spectral analysis and interpretation. A.R., N.R.W. and F.B.B. performed spectral classification. F.B.B. and K.M. correlated the spectra. A.R., D.L.W. and B.S. modelled the light echo. I.T. and D.M. provided imaging of η Car. F.B.B. and D.A.H. provided the FTS images, and F.B.B. and A.R. reduced them.

homework

A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example these contributions).

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

Homework 4:

Data importing and formatting by Victor Ramirez and Liam Kelley.

Astronomy knowledge by Victor Ramirez. Administrative direction by Victor Ramirez

Debugging by Liam Kelley

Mathematical derivation by Shea Fitzgerald

General structure of the code by Shea Fitzgerald

README.md

Partners were Theodore Fessaras, Melvin Tejada, Samuel Matylewicz, and Desi Pilla. We all met up on Saturday to do the lab. I wrote the derivation and data ingestion part on my own the day prior, and compared with the rest of the group when we met up. Desi helped the group a lot with writing the functions. Each of us proofread and error tested his functions and gave feedback.

homework

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

Each student must write a README file for their repository

(20% of the points)

in the readme you must state in your own words

1. what was this homework about? relate it to what we discussed in class
2. what was the hardest part of the homework for you?
3. what was the easiest part of the homework for you?
4. one new thing that you have learned

homework

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

instructions will be here

https://github.com/fedhere/DSP_FBianco

homework

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

online help: bit.ly/3sECCz0
please sign up asap!!

slack

homework

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

of course there is also UD Canvas, which will be used to give you grades and occasionally post messages

midterm

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential. In the Midterm and Final YOU WILL BE EXPECTED TO WORK INDIVIDUALLY.**

- **5% pre-class questions**
- **15% class participation**
- **25% homework**
- **15% midterm**
- **30% final**

Midterm... probably in class

advantages: interviews for jobs are often timed

issues: working under derass is not necessarily a required skill but that is why the midterm counts only 15%!

final

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential. In the Midterm and Final YOU WILL BE EXPECTED TO WORK INDIVIDUALLY.**

- 5% pre-class questions
- 15% class participation
- 25% homework
- 15% midterm
- 30% final

Final: take home, multiple days.

Resources

- SLIDES here
- HOMEWORK INSTRUCTIONS here
- RESOURCES here

If notebooks do not display

use

https://nbviewer.jupyter.org

https://github.com/fedhere/DSPS_FBianco

Resources

Resources

There are several useful textbooks BUT one or even all of them **do not replace learning from lectures and lecture slides.**

The most helpful textbooks are:

- **Statistics, Data Mining, and Machine Learning in Astronomy,** Ivezic, Connoly, VanderPlas, Gray, Princeton Press 2nd edition
- **ML in python: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd edition)** https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ probably the book that is closer to the syllabus in terms of techniques,
- **Python Data Science Handbook,** Jake VanderPlas, O'Reilly Media [https://www.oreilly.com/library/view/python-data-science/9781491912126/] Really specific to the Python Pandas package, but extremely helpful for data science in python

In addition, depending on your familiarity with coding, statistics, and visualization

- computing and coding: **Beginning Python Visualization, 2009**
- **Interactive Data Visualization, S. Murray, O'Reilly Media**
- Visualizations: **Visualizations Analysis and Design, T. Munzer, 2014**

For a more theoretical approach to Machine Learning

- **Elements of Statistical Learning,** Hastie, Tibshirani, Friedman, Springer 2001

Resources

https://www.astroml.org/

Resources

https://github.com/ageron/handson-ml

Resources

Jake Vanderplas is a physicst-data scientists

http://vanderplas.com/

2 what is data science?

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

assume you know

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

neglect

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-sci

PROGRAMMING

STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

PROGRAMMING

STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

python

PROGRAMMING

STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

python

probability
distributions

p-values

uncertainties

MCMC

PROGRAMMING

STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION

http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

python

probability distributions

p-values

uncertainties

MCMC

regression (linear, template)

classification (trees, neural networks)

clustering

2

the scientific method

(what is science?)

3 General principles of "good" science

Falisifiability

Reproducibility

Parsimony

https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

epistemology:

the philosophy of science and of the scientific method

the *demarcation* problem:

what is science? what is not?

My proposal is based upon an *asymmetry* between verifiability and **falsifiability**; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

—Karl Popper, *The Logic of Scientific Discovery*

the *demarcation* problem:

what is science? what is not?

My proposal is based upon an *asymmetry* between verifiability and **falsifiability**; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

—Karl Popper, *The Logic of Scientific Discovery*

a scientific theory must be *falsifiable*

the *demarcation* problem

model → **prediction**

the *demarcation* problem

model → **prediction**

Einstein GR

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}$$

Light rays are deflected by mass

http://discovermagazine.com/2019/may/why-it-took-the-1919-solar-eclipse-for-physicists-to-believe-einstein

the *demarcation* problem

model → **prediction**

data → **does not falsify** → **GR still holds**

position of star changes during eclipse

data → **falsifies** → **GR rejected**

position of star does not change during eclipse

the *demarcation* problem

is phsychology a science?

DISCUSS!

the *demarcation* problem

things can get more complicated though:

most scientific theories are actually based largely on *probabilistic induction* and modern *inductive inference* (Solomonoff, frequentist vs Bayesian methods...)

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

1

assures a result is grounded in evidence

#openscience

#opendata

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

facilitates scientific progress by avoiding the need to duplicate unoriginal research

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

3

facilitate collaboration and teamwork

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

- provide raw data and code to reduce it to all stages needed to get outputs
- provide code to reproduce all figures
- provide code to reproduce all number outcomes

3 the tools

python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

https://www.economist.com/graphic-detail/2018/07/26/python-is-becoming-the-worlds-most-popular-coding-language

python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

https://www.oreilly.com/ideas/2016-data-science-salary-survey-results

python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

https://github.com/fedhere/DSPS2021/blob/master/pythoncrashcourse.ipynb

python crash course by Xiaolong Li https://github.com/xiaolng

a very short introduction to python

basics

- variables and data types number, boolean, string, list, dictionary
- conditionals and loops if, for, while
- functions
- classes
- standard library

packages

- numpy/scipy numerical/scientific computing
- pandas data analysis and manipulation
- matplotlib visualization
- statsmodels statistical modeling
- sklearn machine learning
- keras/tensorflow pytorch deep learning

resources

python

https://sharmamohit.com/work/courses/ucsl/

series of notebooks designed for Urban Science students by Dr. Mohit Sharma (in consultation with me)

recommanded if you are brand new to python and coding or are serious about cleaning up your foundamentals

Overview

Welcome

Testing Welcome

Accessing Jupyter Notebook

Instructions on accessing jupyterhub

Introduction to Notebook and beyond

Introduction to features of Notebooks and python syntax

Introduction to Python

Description of Module 1

Python-Variables

python

quick bootcamp

will set it up in a couple of weeks

recommanded if you know some python or if you know some other conding language reasonably proficiently

https://github.com/fedhere/PyBOOT

Table of Contents

1 Native variable types
1.1 strings, int, floats
1.1.1 print formatting
1.2 bool
1.2.1 if/else statements with bools
1.2.2 concatenating bool statements
1.2.3 math with bools
1.3 lists
1.4 dictionaries
2 IDE other than jupyter notebooks
2.1 python
2.2 ipython
2.3 execute python from the shell
3 Numpy types
4 numpy arrays
5 PART 2: Slicing, Broadcasting, and math operators on arrays and lists
5.1 operations with arrays
5.2 slicing
6 PART 3: Functions
7 file IO
8 PART 4: multi dimensional arrays
9 Part 5: iterators - for loops, enumerate, and list comprehensions
9.1 for loops
9.2 enumerate
9.3 list comprehension
10 PART 6: matplotlib
10.1 setting up pylab plotting
10.2 figures and axis objects and simple plots
10.3 plotting errorbars
10.4 plotting 2D arrays


```
from __future__ import print_function, division
#importing this to make code python2&3 compatible
# overwrites the default print to require parenthesis
# overwrites the default / (division operator) behavior
#   so division of 2 integers returns a float
```

python

online book

https://www.southampton.ac.uk/~fangohr/training/python/pdfs/Python-for-Computational-Science-and-Engineering.pdf

Introduction to

Python for Computational Science and Engineering

(A beginner's guide)

Hans Fangohr
Faculty of Engineering and the Environment
University of Southampton

September 7, 2015

python

PEP8: Python Enhancement Proposals 8

"This document gives coding conventions for the Python code comprising the standard library in the main Python distribution."

Indentation, *Tabs vs Spaces*, *Maximum Line Length*, *Blank Lines*, *Source File Encoding*, *Imports*, *Whitespace in Expressions and Statements* , *Imports*, *Comments Bookeeping*, *Naming*

github

reproducibility

https://github.com

Reproducible research means:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

allows reproducibility through code distribution

github

version control

https://github.com

the Git software

is a distributed *version control system:*

a version of the files on your local computer is made also available at a central server.

The history of the files is saved remotely so that any version (that was checked in) is retrievable.

allows version control

github

collaborative platform

https://github.com

collaboration tool

by fork, fork and pull request, or by working directly as a collaborator

allows effective collaboration

Jupyter Notebook
Google Colaboratory

https://colab.research.google.com/notebooks/welcome.ipynb#

stackoverflow

for when you need help

https://stackoverflow.com/

you can ask coding questions, installation questions, colab questions...

How to type list comprehensions

Ask Question

0

I have the following list comprehensions in Python:

```
from typing import cast

# everything is fine
print([value for value in [1, 2, 3, 4]])

# on the first "value": Expression type contains "Any" (has type "List[Any]")
print("{}".format([value for value in [1, 2, 3, 4]]))

# on the "cast": Expression type contains "Any" (has type "List[Any]")
print("{}".format([cast(int, value) for value in [1, 2, 3, 4]]))
```

Why does using `format` cause Mypy to give me back errors? As you can see, I tried to use casting and it still failed.

This question looks similar, but my particular case is weird because Mypy seems to be fine as long as I'm not using the `format` function (yet it's always okay with the `print` function).

Is there anything I can do to not have the lines with formatting give me errors? (Or should I just `# type: ignore` them?)

python | python-3.x | list-comprehension | typing | mypy

stackoverflow

for when you need help

https://stackoverflow.com/

you can ask coding questions, installation questions, colab questions...

Multiple output regression or classifier with one (or more) parameters with Python

Ask Question

5

I wrote a simple linear regression and decision tree classifier code with Python's Scikit-learn library for predicting the outcome. It works good.

My question is, Is there a way to do this backwards, to predict the best combination of parameter values based on imputed outcome (parameters where accuracy will be the best).

Or I can ask like this, is there a classification, regression or some other type of algorithm (decision tree, SVM, KNN, logistic regression, linear regression, polynomial regression...) that can predict multiple outcomes based on one (or more) parameter/s?

I have tried to do this with putting multivariate outcome, but it shows the error:

```
ValueError: Expected 2D array, got 1D array instead:
array=[101 905 182 268 646 624 465].
Reshape your data either using array.reshape(-1, 1) if your data has a single feature
```

This is the code that I wrote for regression:

```
import pandas as pd
from sklearn import linear_model
from sklearn import tree

dic = {'par_1': [10, 30, 13, 19, 25, 33, 23],
       'par_2': [1, 3, 1, 2, 3, 3, 2],
       'outcome': [101, 905, 182, 268, 646, 624, 465]}
```

stackoverflow

for when you need help

it can be a toxic environment...

https://stackoverflow.com/

you can ask ciding questions, installation questions, colab questions...

Multiple output regression or classifier with one (or more) parameters with Python

Ask Question

5

I wrote a simple linear regression and decision tree classifier code with Python's Scikit-learn library for predicting the outcome. It works good.

My question is, Is there a way to do this backwards, to predict the best combination of parameter values based on imputed outcome (parameters where accuracy will be the best).

Or I can ask like this, is there a classification, regression or some other type of algorithm (decision tree, SVM, KNN, logistic regression, linear regression, polynomial regression...) that can predict multiple outcomes based on one (or more) parameter/s?

I have tried to do this with putting multivariate outcome, but it shows the error:

```
ValueError: Expected 2D array, got 1D array instead:
array=[101 905 182 268 646 624 465].
Reshape your data either using array.reshape(-1, 1) if your data has a single feature
```

This is the code that I wrote for regression:

```
import pandas as pd
from sklearn import linear_model
from sklearn import tree

dic = {'par_1': [10, 30, 13, 19, 25, 33, 23],
       'par_2': [1, 3, 1, 2, 3, 3, 2],
       'outcome': [101, 905, 182, 268, 646, 624, 465]}
```

key concepts

What is Data Science

Reproducibility

1

homework

- make an account on GitHub if you do not have one yet
- Create a repository called DSPS_<firstinitialLastname>

answer this form and share your github repository link in it
https://forms.gle/GtufqNMfF9gCnzoG6

there is also a longer exercise that you can do if you want to learn more github tricks. This is not mandatory tho

https://github.com/fedhere/DSPS_FBianco/tree/master/lab1

Jeff Leek & Rodger Peng.
2015,
What is the Question?

http://fbb.space/dsps/The%20Research%20Question-2015-Leek-1314-5.pdf

reading

the original link:
https://science.sciencemag.org/content/347/6228/1314.summary
is link nees access to science magazine, but ou can use the link above which is the same file)

STATISTICS

What is the question?

Mistaking the type of question being considered is the most common error in data analysis

By **Jeffery T. Leek** *and* **Roger D. Peng**

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

http://sepwww.stanford.edu/data/media/public/docs/sep67/jon2/paper_html/

Reproducibility and Replicability in Science

Consensus Study Report (2019)

https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

additional reading