Data Science with Python CS677

Farshid Alizadeh-Shabdiz, PhD, MBA
Alizadeh@bu.edu
Fall 2021

1

High Level View

2

Identify the risk factors of prostate cancer

- 97 patients collected by Stanford's physician
- Scatter plot matrix of pair of variables

3

FiveThirtyEight – Started as New York Times blog

- An American website that focuses on opinion poll analysis, politics, economics, and sports blogging.

- Successfully predicted 2012 Senate and presidential outcome.
- Got acquired by ESPN and later by ABC.

4

Statistical Learning Problems
Identify the risk factors of heart disease

- Framingham heart study (FHS)

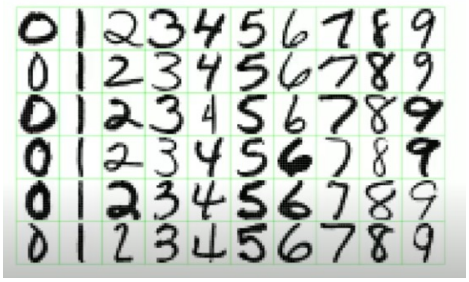
"FHS **findings** have informed the understanding of how cardiovascular health affects the rest of the body. The **study** found high blood pressure and high blood cholesterol to be **major risk** factors for cardiovascular disease."

- Initial number of patients 5200
- Resulted to a risk score to estimate 10-year cardiovascular risk, showed risk factors as
  - Age, total cholesterol, smoking, and systolic blood pressure
  - Very small correlation with HDL cholesterol

5

Identify handwritten numbers

- Example is MNIST dataset with 60,000 samples.

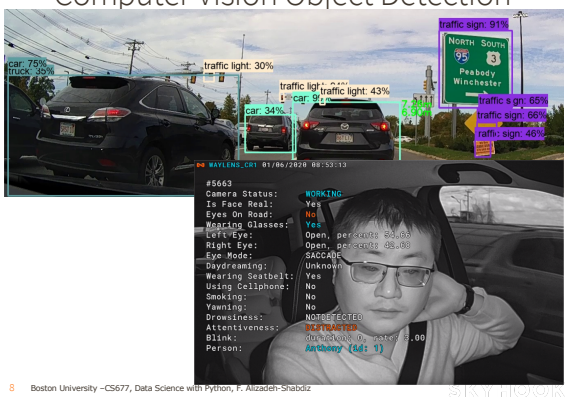


6

- Image segmentation
- Object detection

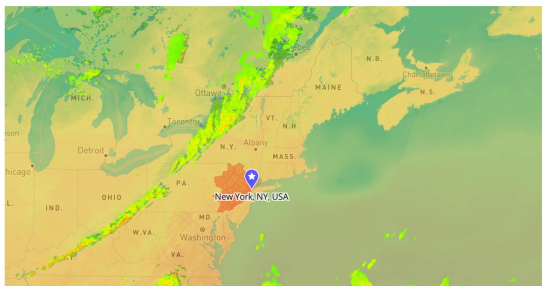
7

Computer Vision Object Detection



8

Weather Prediction?



9

Wi-Fi Localization and GNSS

10

Examples of Big Data Science Markets

- Autonomous vehicles
- Weather prediction
- Recommendation engine
- Add industry - user segmentation or extracting personas of users

11

What is Data Science?

Learning from data

Summarizing data

Presenting data

Modeling data (precision and uncertainty) + Inference

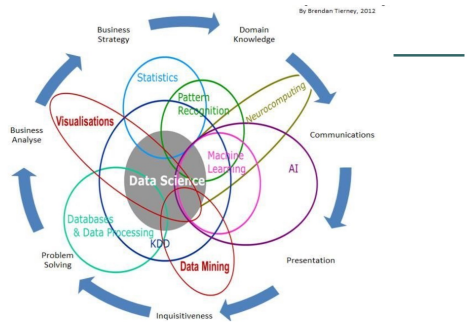
12

Data Science Outcome

- Predicting high risk heart patients
- Finding risk factor in heart attack
- Finding important factors in heart attack
- Finding important factors in heart attack in different races

13

Data Science Multidisciplinary



14

Steps of Data Analysis

- Data collection
- Data warehousing/data preparation
- Data preparation
  - Data cleansing
  - Outlier removal
  - Missing data
  - Duplicates
- Making sense of data – e.g. statistics, visualization
- Data science algorithm(s)
- Commercialization - big data technologies (Hadoop, Spark, etc.)

15

Data Science

- Two big classes of problems
  - Regression
  - Classification
- Three groups of algorithms
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

16

Simple Models

17

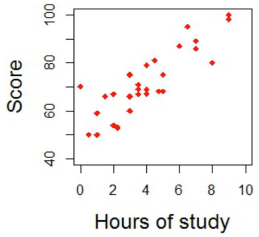
Linear Regression - Simplest Regression Model

- Linear regression

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- Optimization criteria

$$E[(Y - f(X))^2 | X = x]$$



18

Generalized Linear Model

- Linear regression

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- General linear

$$f(X) = ag(x) + b$$

Example of *g(x):* polynomial

19

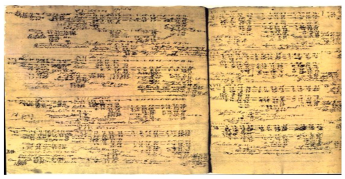
General Algorithmic Approach

- Choose general model(s)
- Collect data
- Clean data
- Divide data to training and testing set
- Train model(s) using the training set
- Choose the best model using the testing set, including selecting *hyperparameters*

20

Example: Kepler's Laws

- Johannes Kepler worked for Tycho Brahe
- Brahe compiled detailed observations (especially Mars)



Reference: Prof Pinsky's lecture.

21

Distance, Periods Data

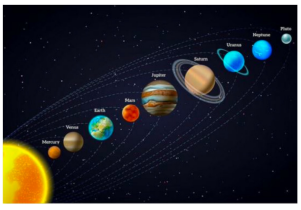
- Kepler "cleaned" data:

Planet	Period (days) T	Distance (AU) to Sun R
Mercury	87.77	0.389
Venus	224.70	0.724
Earth	365.25	1
Mars	686.95	1.524
Jupiter	4332.62	5.2
Saturn	10759.2	9.510

- what is R vs. T?
- Kepler discovered $R^3 = aT^2$

22

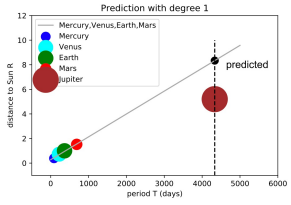
Periods and Orbits



"I first believed I was dreaming
But it is absolutely certain and exact
that the ratio which exists between
the period times of any two planets
is precisely the ratio of the 3/2th
power of the mean distance."
translated from Harmonies of the World by Kepler (1619)"

23

Kepler's Question

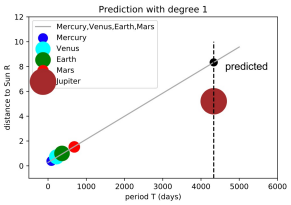


$$f(R) = ag(T) + b$$

- how are R and T related?
- what link functions $f(R)$ and $g(T)$ match the data?

24

Linear Prediction

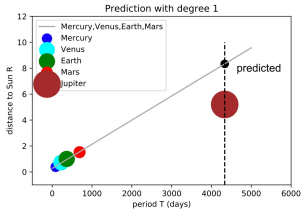


$$R = aT + b$$

- training set: Mercury, Venus, Earth, Mars
- testing set: Jupiter

25

Linear Prediction

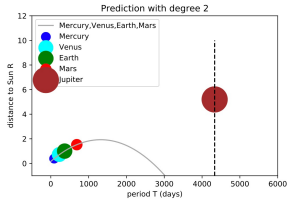


$$R = (1.86 \times 10^{-3}) \cdot T + 0.27$$

- predicted $R(\text{Jupiter}) = 8.33$
- "exact": 5.2 (rel. error 60%)

26

Quadratic Prediction

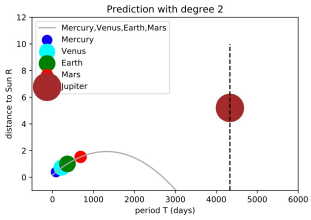


$$R = aT^2 + bT + c$$

- polynomial link for $g(T)$
- training set: Mercury, Venus, Earth, Mars

27

Quadratic Prediction

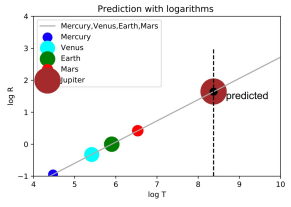


$$R = -(1.01 \times 10^{-6}) \cdot T^2 + (2.67 \times 10^{-3}) \cdot T + 0.17$$

- predicted $R(\text{Jupiter}) < 0$!!!

28

log-log Prediction

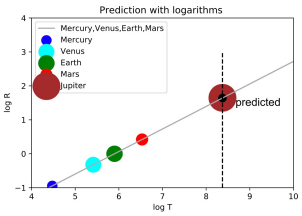


$$\log R = a \log T + b$$

- logarithmic link for f and g
- log() function was invented by John Napier in 1614

29

log-log Prediction



$$\log R = 0.66 \cdot \log T - 3.91$$
$$3 \cdot \log R = 2 \cdot \log T - 11.73$$
$$\Rightarrow R^3 \propto T^2$$

30

Kepler Result

Planet	Period (days) T	Distance (AU) to Sun R	R^3/T^2 $10^{-6}\text{AU}^3/\text{day}^2$
Mercury	87.77	0.389	7.64
Venus	224.70	0.724	7.52
Earth	365.25	1	7.50
Mars	686.95	1.524	7.50
Jupiter	4332.62	5.2	7.49
Saturn	10759.2	9.510	7.43

- Kepler's third law $T^2 \propto R^3$

square of orbital period of a planet is proportional to the cube of the semi-major axis of its orbit

31

Modern Data

Planet	Period (days) T	Distance to Sun R	R^3/T^2 $10^{-6}\text{AU}^3/\text{day}^2$
Mercury	87.9693	0.38710	7.496
Venus	224.7008	0.72333	7.496
Earth	365.2564	1	7.496
Mars	686.9796	1.52366	7.495
Jupiter	4332.8201	5.20336	7.504
Saturn	10775.599	9.53707	7.498
Uranus	30687.153	19.1913	7.506
Neptune	60190.03	30.0690	7.504

- "close" match to Kepler's
- Earth: 7.5 vs. 7.496

32

Review of the Past

- Correlation
  - Properties
  - One and minus one meaning
- Confidence interval
- P-value
- Hypothesis testing
- Central limit theorem
- Linear regression
  - Explanatory input / independent variable
  - Response/output/dependent variable
  - Intercept and slope
  - Categorical input
  - Outlier – influence point
  - Co-linearity - Correlated explanatory inputs
  - Causation vs association
  - Assessment – MSE
  - Assessment of goodness of a model – R2 / Adjusted R2 / F-test
- Multi linear regression
- Polynomial regression

33

Regression Assessment

- Fitness functions still can be used
- Or family of assessment like
  - R squared

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}}$$

  - Adjusted R squared

$$R^2_{adj} = 1 - \left(\frac{Res\ SS/(n-k-1)}{Total\ SS/(n-1)}\right)$$

  - F-test

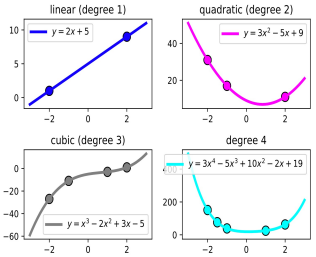
$$F = \frac{Reg\ SS/Reg\ df}{Res\ SS/Original\ model\ df}$$

Reference: Wikipedia

34

Generalized Linear Model

- Overfitting / Underfitting
- Interpretability / Complexity (Occam's Razor [OK] + [UHMZ] + [RAY] + [ZUH])



35

Logistic Regression

36

Ex: Logistic Regression

- assume probability P
- define *odds* as

$$\text{odds}(P) = \frac{P}{1-P}$$

- $P = 0.25 \mapsto \text{odds}(P) = 1/3$
- $P = 0.50 \mapsto \text{odds}(P) = 1$
- $P = 0.75 \mapsto \text{odds}(P) = 3/1$
- want to model

$$f(\text{odds}(P)) = ag(x) + b$$

37

Main Idea

- use $f = \log()$ as *link* for the *odds*
- use regression

$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

$$\frac{P}{1-P} = \exp(b_0 + b_1 x)$$

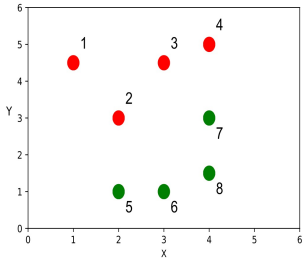
$$P = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$$

- note:

$$\frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)} = \frac{1}{1 + \exp(-(b_0 + b_1 x))}$$

38

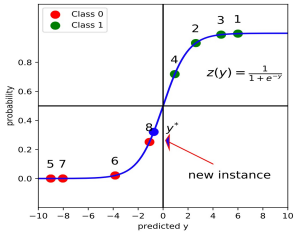
Original Dataset



- use *logit*() function

39

Computing Class Labels



- $z(y^*) < 0.5$ - "red" (class 0)

40

Summary

- linear functions for predictions and/or classification are simple and easy to explain
- most models are not linear
- idea: use *link* functions to transform models
- look for linear functions in the transformed space
- use these functions to solve our original problem

41

Data Science with Python CS677 – What is Covered?

42

Philosophy of the Course

1. Learning data analytics packages/libraries of Python
2. Learn classifier techniques
   - Helps to apply them correctly – know how and when to use them
   - Assess algorithms
   - Be able to choose - Ocham's razor principle
3. Learn by doing – combine 1 and 2

43

Data Science CS677 What is the course about?

- Learning Python in the context of data science
- Learning by doing!
  - Learning about some of the fundamental data science algorithms in depth
  - Use python to implement them – will be using python libraries

44

What is covered under Python?

- Review of python
- Numpy
- Pandas
- Data Visualization
- Data Wrangling, Aggregation

45

Machine Learning Aspect of the Course

- Statistical machine learning
  - Statistical and predictive analysis
  - Linear regression
  - Logistic regression
  - Linear discriminant analysis
  - Decision trees
  - Naïve Bayes
  - K-Nearest Neighbors
  - Support Vector Machines, etc.

46

Course Admin & Environment and Software Setup

47

Admin

- Pre-requisite: knowing Python (CS 521)
- All the material will be posted on Blackboard
- There will be quizzes every couple of weeks - online
- Assignments every week, which are due before the next class
  - No late assignment will be accepted
- Office hours: by appointment
- TAs intro – TAs will also announce their office hour

48

Course Material

- Course material
  - Notes and material are the main source
  - Text book is a good additional resource
- Academic code of conduct

49

Grading

- Grading
  - Assignments: 30%
  - Quizzes – 25%
  - Term project – 15%
    - To apply what we learn yourself - **PLUS** additionally you can apply a new algorithm
    - A data science project using python
    - You have to choose a dataset
    - Projects are individual
  - Final exam – 30%
    - From the entire class

50

Python

51

Why Python?

- Python became de-facto data science language
- Power of general-purpose programming languages
- Ease of use
- Libraries for everything
  - data loading, visualization, statistics, natural language processing, image processing, and more.
- Ability to interact directly with the code, using a terminal or tools like the Jupyter Notebook
- Allow quick iteration and easy interaction.
- Large active user community

52

Python

- Python is a scripting language
- Python is an interpretive language
- You don't need to define the data type
- Two groups of data types
  - Primitive: Boolean, Integer, Flout, complex, Char
  - Collection: Set, List, dictionary, Tuples, Strings

53

- Scikit-learn
  - Very active user community
  - Contains a number of state-of-the-art machine learning algorithms
  - Comprehensive documentation
  - The most prominent Python library for machine learning.

54

Jupyter Notebook

- An interactive environment for running code in the browser.
- Great tool for exploratory data analysis
- Widely used by data scientists
- The Jupyter Notebook makes it easy to incorporate code, text, and images

55

numpy

- NumPy is one of the fundamental packages for scientific computing in Python.
- It contains functionality for multidimensional arrays, high-level mathematical functions such as linear algebra operations.
- The core functionality of NumPy is the ndarray class, a multidimensional (*n*-dimensional) array.
- All elements of the array must be of the same type.

56

SciPy

- SciPy provides functions for scientific computing
- It provides advanced linear algebra routines, mathematical function optimization, signal processing, special mathematical functions, and statistical distributions.
- scikit-learn draws from SciPy's
- SciPy has scipy.sparse, which provides *sparse matrices*, which are another representation that is used for data in scikit- learn. Sparse matrices are used whenever a 2D array that contains mostly zeros:

57

pandas

- pandas is a Python library for data wrangling and analysis.
- DataFrame (that is modeled after the R DataFrame) is the center of the design
- DataFrame is a table that can house different kind of data
- Only columns have to be the same data type

58

matplotlib

- Most important plotting library in Python
  - line charts, histograms, scatter plots, and so on.
- Inside the Jupyter Notebook, you can show figures directly in the browser

59

Installation

- Anaconda
  - A Python distribution made for large-scale data processing, predictive analytics, and scientific computing.
- *Enthought Canopy*
  - Another Python distribution, which comes with NumPy, SciPy, matplotlib, pandas, and IPython, but the free version does not come with scikit-learn.

60

Python 2 vs 3

- Major versions of Python
  - Python 2 version 2.7
  - Python 3
- Python 3 was a major upgrade to Python 2
- Python 2 is no longer actively developed
- Python 2 code usually does not run on Python 3
- We always use Python 3
  - Unless there is a major python 2 legacy code

61

Software Setup

- Python 3 (3.x version)
  - https://www.python.org/downloads/
- Install Jupyter
  - python3 -m pip install --upgrade pip
  - python3 -m pip install jupyter
- For MACs
  - /Applications/Python 3.8/Install\ Certificates.command
- Run Jupyter
  - jupyter notebook

62

Install Setup

- Install Python Modules by using "python3 –m pip install <x>
- Packages:
  - Data manipulation: Pillow , requests
  - Analytics packages: numpy, pandas
  - Visualization: matplotlib, seaborn, plotly
  - Machine learning: scipy, scikit-learn, statsmodels

63

Python

- Load the code with "ipynb" extension
- Under "kernel" at the top menu, you can select "Restart & Clear Output"

64

Python basic commands

- def function name:
- Data types
  - String, tuples, sets
  - Dictionary, list

65