



DATA SCIENCE

PPGla/PUCPR

Prof. Jean Paul Barddal



About

- Jean Paul Barddal
- Data Stream Mining
- jean.barddal@ppgia.pucpr.br
- www.jpbarddal.com.br
- Research topics: Machine learning
(classification, regression, clustering, feature selection, recommender systems) for streaming data
- Applied ML: Financial systems, Education, Recommender systems for e-commerce, Log analysis, etc



Warning

- We are in a Level I Global Classes program
- This means that:
 - Content (slides, activities, test, etc) will be in English
 - We will talk in Portuguese
 - The test will be in English, but you may answer them in either Portuguese or English

Warning

- This is **not** a crash course on Data Science using Python
- You are highly expected to take your time to learn more about the tools we will use (numpy, pandas, scikit-learn, etc)
- We are interested in both your coding skills and also in your critical reasoning

Agenda

- **March 1st** - Lecture 1 – Overview, grading, Basic Statistics
- **March 8th** - Lecture 2 – Univariate data analysis
- **March 15th** - Lecture 3 – Multivariate data analysis
- **March 22nd** - Lecture 4 – Correlations
- **March 29th** - Lecture 5 – Enhanced data visualization
- **April 5th** – Lecture 6 – Missing data & outliers
- **April 12th** - Lecture 7 – PCA and t-SNE
- **April 19th** - Lecture 8 – Test

Grading

- We have 8 meetings and you must be present in 75% of them, i.e., 6 lectures
- Your grading will be based on a test to be done on April 19th (A \geq 9, B \geq 8, C \geq 7, D otherwise)

Slides

- Slides will be made available on my website
- www.jpbarddal.com.br

Recordings

- Lectures will **NOT** be recorded
- There are no **IFs** and no **BUTs** on this

Polls

- You need to attend the poll that will be made in the beginning of each lecture
- Throughout the lecture, if you are requested to participate, and you do not, you may be assumed as absent

Time to introduce yourself

- I will now invite each one of you to introduce yourself, along:
- Affiliation
- Degree you intend to acquire (msc/phd)
- Institution
- Advisor
- Project description
- Your expectations for this discipline

ENVIRONMENT SETUP

Google Colaborate

- Hereafter we will use Google Colaborate
- It will allow us to run Python code in the cloud
- Most part of the data analytics and machine learning tools are available there

Set up your account now :)

<https://colab.research.google.com/>



Anaconda

- If you're not too keen on working on the cloud, you should be able to use Jupyter and (preferably) Anaconda
- Anaconda allows you to keep different Python versions, each with different packages

<https://www.anaconda.com/>



PANDAS

Pandas

- Pandas is the most popular and good tool for handling data and data analysis
- Let's focus on tabular data for now



Tabular data

Columns are called:
Attributes, Features
Variables, Fields, Characteristics

Lines are called:
Objects
Instances
Samples
Registers
Cases

Gender	Age	Salary	Job Role	Married?
Male	28	5600,34	Programmer	N
Male	22	3215,50	Data Analyst	Y
Female	32	12000,00	Project Manager	N
Female	27	4500,00	Lawyer	N
Male	17	1400,00	Accounting Intern	N
...

Types of variables

- Numeric
 - Interval
 - Money quantity, temperature in Celsius, Fahrenheit, etc
 - Ratio
 - The same as above, yet, 0 has a special meaning
 - Height, weight, temperature in Kelvin (note that these cannot be negative!)
- Categorical
 - Nominal
 - Ex.: Gender (M/F), Nationality, Car make
 - Ordinal
 - Ex.: Number of stars (hotel rating), movie ratings (poor, good, great)

Descriptive analytics

- Goal: summarize and describe a dataset
- Main goals:
 - Minimum and maximum values
 - Mean
 - Median
 - Mode
 - Variance
 - Standard deviation

Mean

- Sum of all values divided by the amount of values

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mode

- The most repeated value in data
- What is the mode in each of the lists below?

• [1, 1, 2, 3, 4] $\{1\}$

• [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] *Amodal*

• [1, 1, 1, 2, 2, 2, 3, 4, 5] $\{1, 2\}$ *Bi modal*

• [1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5]

$\{1, 2, 3, 4, 5\}$ *5-modal*

$f \geq 2$

Median

- Given a sorted dataset, the median is the value that is in its center position
- Example:

[32, 33, 24, 31, 44, 65, 32, 21, 32]

Sorting:

→ [21, 24, 31, 32, 32, 32, 33, 44, 65]

Median:

— 50% — | — 50% —
↓

[21, 24, 31, 32, 32, 32, 33, 44, 65]

Median

- If the amount of values is even, the median is given by the average of the center positions
- Example:

→ [18, 19, 19, 22, 44, 45, 46, 46, 47, 48]

Median = $(44+45)/2 = 44,5$

Quartiles and Percentiles

- Quartiles divide the data in 4 parts. These are indicated by Q1, Q2 and Q3, such that Q2 is the median

Q1 = 19 Q3 = 46,5

idades = [18, 19, 19, 22, 44, 45, 46, 46, 47, 48]
(anos)

Q2 = 44,5

$$\begin{aligned}\bar{x} &= 35,4 \text{ (anos)} \\ \sigma^2 &= 170,44 \text{ (anos}^2\text{)} \\ \sigma &= 13,06\end{aligned}$$

- The same rationale can be applied to percentiles, that divide the data in each 1%: P1, P2, P3, ... P99

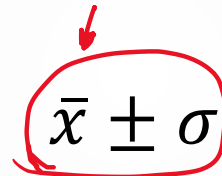
Variance

- Given a dataset, the variance tells us how distant each value is from the mean
- The smaller the variance, the closer all values are from the mean
- Variance is given by:

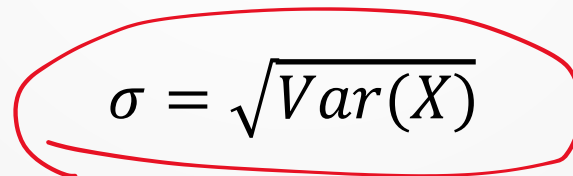
$$\boxed{Var(X)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Standard deviation

- The standard deviation tells us the "error" in a dataset if a value is replaced by the mean
- The standard deviation is often showed next to the mean:


$$\bar{x} \pm \sigma$$

- And it is the square root of the variance

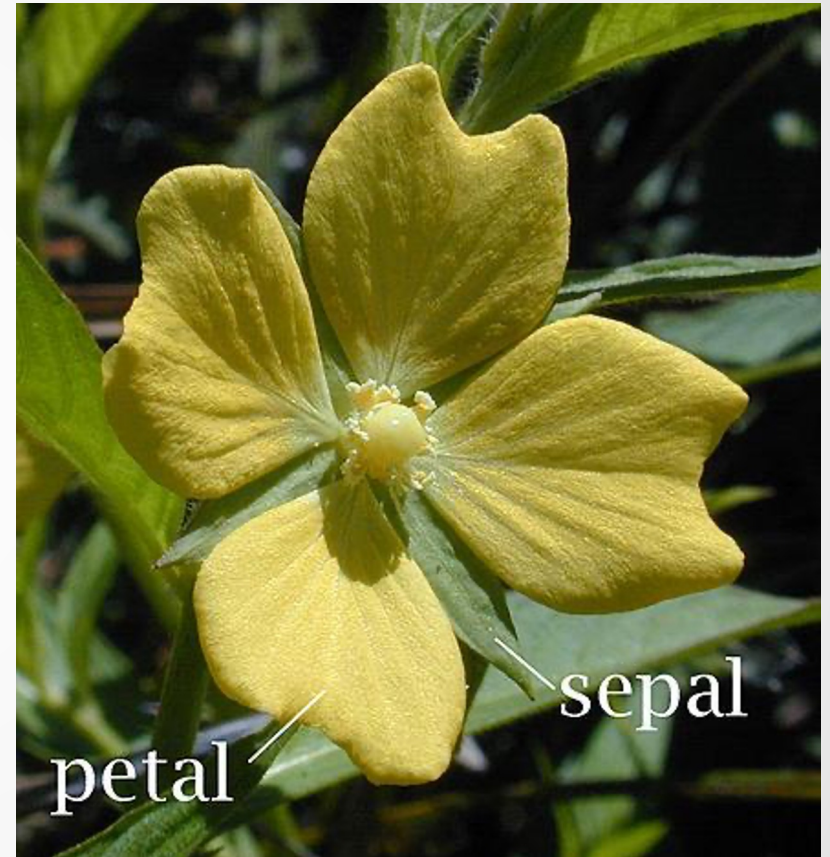

$$\sigma = \sqrt{\text{Var}(X)}$$

LET'S HAVE A BREAK: 20 min!

ACTIVITY

Activity 1 - Using the Iris Dataset

- Download the notebook
- Perform all the operations in it with the iris dataset
- Attributes
 - Petal Length
 - Petal Width
 - Sepal Length
 - Sepal Width



Activity 2

- Assemble in pairs!
- Each pair will receive a specific dataset
- You should follow the link below and replace X with your team's number.
- <https://jpbarddal.github.io/assets/data/datascience/ans/datasetX.csv>

Activity 2 – let's continue

- Now, two teams should unite and discuss their findings
- What are the main statistics you have computed?
- What do you think is going on with these datasets?

WHAT IS GOING ON?

Anscombe Quartet

Same Stats, Different Graphs:

Generating Datasets with Varied Appearance and Identical
Statistics through Simulated Annealing

Anscombe Quartet

