Weekly meeting 2

Dr. Doina Bein

Tuesday, June 13, 8:30-10am

Surveys to be completed

To be done today, before starting research:

CIC-PCUBED Pre-event survey:

https://fullerton.qualtrics.com/jfe/form/SV_6YIVSkC6hLxbunA

Project 1: Data Science

What you need to do: topics & objectives

Objective 1: Learn Python using some textbook or some online courses such as (https://www.codecademy.com/learn/learn-python). Shared by Stephanie Pocci: Learn Python in a couple hours. This YouTuber does a very beginner-friendly crash course about the capabilities of Python and its uses. Here is the link: https://www.youtube.com/watch?v=rfscVS0vtbw

Objective 2: Learn how to use Jupyter Notebook. Start here http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html

Objective 3: For data science, find a suitable dataset and start training some neural network using with Google tensorflow.

Logistics for all students

- Who is participating: list of current research students and their availability
- Research will be conducted virtually during the week with in-person meetings throughout the week
- Zoom meetings for me to teach new topics and for you to participate in open discussions
- Support:
  - If needed, you can meet me
    Zoom: Mon, Tu, Wed from 8:30-10:25 am
    IN PERSON: Mon, Tu, Wed from 8:30-9:30 am, Thursday 8:30-10am or by email
  - CIC-PCUBED peer mentor: (tentative) availability

Logistics for all students (contd.)

- Make a copy of this GDoc Work schedule, share the Gdoc copy with me, and maintain it weekly and daily; due at the end of Week 2
- Before the end of week 3, make a copy and maintain your Proposed work by individual or teams of up to three; due by the end of Week 3
- Complete your availability here; try to have it consistent over the 7 weeks such that it will be easy to partner in the project
- Group projects: to be decided; sample list here
- Oral or poster presentations: tentatively scheduled for Friday, July 28, from 8:30am-12:30 pm and if needed, from 1:30-4 pm

Please checkout:

- Other websites and ebooks
- Websites with free datasets
- If you find good, free resources, please share it by email or during weekly meetings
- Next meeting: I will lecture on ZOOM on Data Science: Thursday, June 15, from 10:30am-12pm

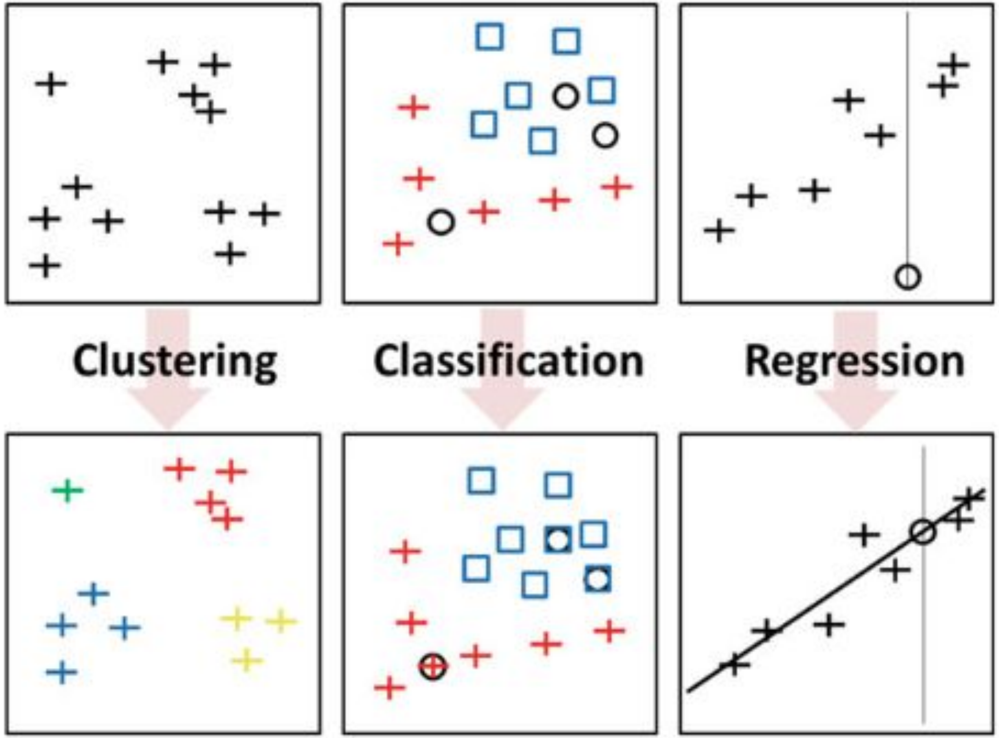
Progress on Learning Python

- Free course: https://www.codecademy.com/learn/learn-python
- Free course: https://www.kaggle.com/learn/python
- Youtube video (about 4 hours): https://www.youtube.com/watch?v=rfscVS0vtbw

Data Science

Linear Algebra in Data Science

- *Data Science* (DS) heavily relies on efficient implementation of linear algebra primitives in these three main categories:
- *Clustering*. We seek for homogeneous group of data in data-sets: It is a class discovery procedure in data exploratory, also called *unsupervised classification*.
- *Classification*. Given a training set of data labeled with their class, we seek to label new unlabeled data by means of a *classifier*. That is, we predict a discrete class variable. It is also called *supervised classification*.
- *Regression*. Given a data-set and a function on this data-set, we ask for the best model of the function that explains the data. This allows one to interpolate or extrapolate values of this function for new data. In general, *regression* is a mechanism that allows to study the relationship of a variable with another.



Clustering

Classification

Regression

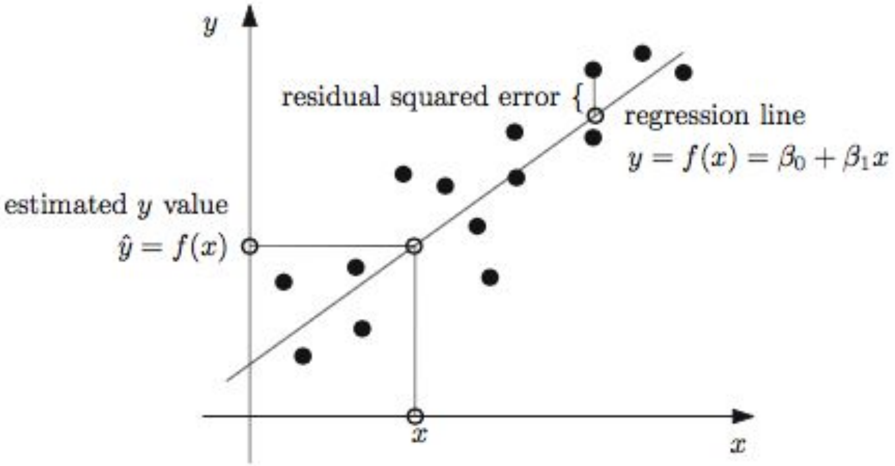
Linear Regression Modeling of Data-sets

- Principle of the linear regression modeling of data-sets: we are asked to predict the value $\hat{y} = f(x)$ of a function at a query position x with: $f(x) = \hat{\beta}_0 + \sum_{i=1}^{d} \hat{\beta}_i x_i$
- We augment the dimensionality of data by adding an extra coordinate $x_0 = 1$ to unify the function evaluation as a *dot product:*
  - consider $x \leftarrow (x, 1)$
  - and $f(x) = \sum_{i=0}^{d} \hat{\beta}_i x_i = x_i^T \beta$
  - We are given a collection of observations $\{(x_1, y_1),\ldots,(x_n, y_n)\} \in R^d$ and we want to fit the best model function by minimizing the *Residual Sum of Squares* (*RSS*):

$$\hat{\beta} = \min_{\beta} \sum_{i=0}^{d} (y_i - x_i^T \beta)$$

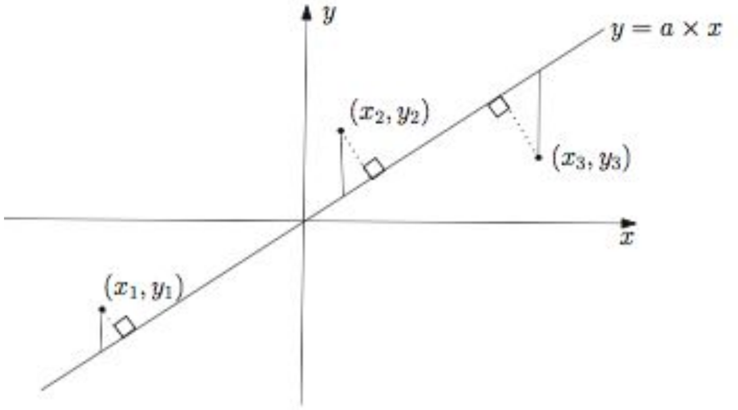
Ordinary Linear Regression

- The ordinary linear regression considers a data matrix *X* of dimension *n* × (*d* + 1) with a column vector *y* of dimension *n* and the hyperplane parameter vector β of estimates of dimension *d* + 1



Total Linear Regression

- A different approach would be to measure the error as the squared *orthogonal projection* lengths of data to the predicted values: this is called the *total regression* or *total least squares method*



Total regression (total least squares) seeks to minimize the squared orthogonal projection lengths of data to the model hyperplane.

Ordinary regression (ordinary least squares) asks to minimizes the squared vertical projection lengths of data to the model hyperplane.

- Total least squares is more complicated to calculate as there is no simple closed-form solution for solving it directly.
- Regression can also be used for classification

Fun facts about yourself

- One fun fact about yourself
- Round table discussions:
  - What topics in CS you see them needed in the future?
  - What do you think about learning Python? Any cool feature you want to share?