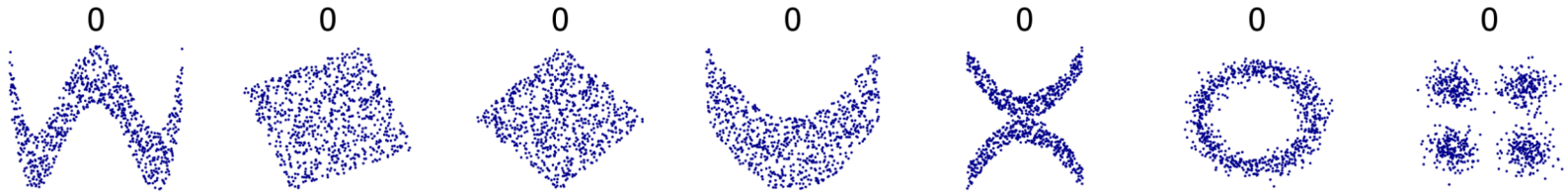
Tutorial: Machine Learning in Precision Medicine
SS 2019

Jana Fehr (jana.fehr@hpi.de)
Remo Monti (remo.monti@guest.hpi.de)

Exercise 1: Linear Regression

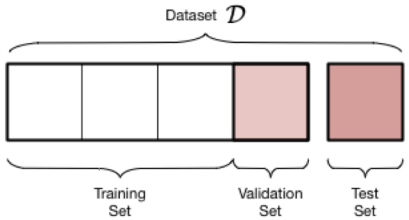
- Question 1.1: Are two variables *x, y* always independent if their covariance is 0?



- Covariance difference between calculations with lists and numpy:
  - → Vanilla: biased numerator n
  - → Numpy: unbiased with n-1 numerator
  - → since we are looking at sample observations and not at the whole population.

Exercise 2: Linear Regularization

- PIMA Indian Dataset on diabetes
- Multiple variables to predict Insulin level
- Experiment Design:



- Regularization
  - Why needed?
  - Mathematic matrix notation

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}$$

$$\boldsymbol{\theta} = \left[\,\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda I\right]^{-1}\left[\,\frac{1}{n}\mathbf{X}^T\mathbf{Y}\,\right]$$

Goals of this assignment

- Get familiar with pandas functions for pre-processing
- Get familiar with numpy matrix-operations
- Understand the concept of designing a machine learning experiment (train-, valid- and test sets)
  - scikit-learn
  - Hyper-parameter tuning
- Understand the concept of Regularization.

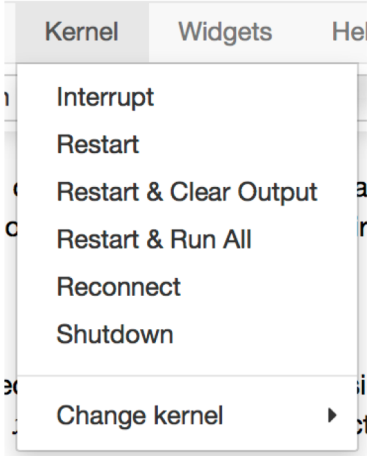
For future assignments

- Run every cell of the notebook, even if the cell does not require your code
- At the end of your exercise: Restart kernel and run from the beginning to make sure that everything works.
- Send the .ipynb file only. No zip, no .dms, or .txt files
- Ask questions in eMail, not notebook
- Groups are not fixed, you can still form groups.
- Mailing list: ml_precmed_sose2019

  Is everyone signed up?

- Submission process:

  MOOC? Open HPI experience?



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank you
for your attention!

Remo & Jana