McGill **Artificial Intelligence** Society

Intro to Data Preprocessing

By Li Zhang and Diego Lopez

Follow Along

Slides @ tiny.cc/mais-f2019-workshops

Agenda

- Why Python?
- Overview of Python syntax
- Some motivation for preprocessing
- Hands-on demo

Why Python?

- Open source libraries
- “Simple is better than complex”
- Python Notebooks
- Industry standard

Python Syntax

Assignments

```
my_num = 4
my_string = “hello”
# Can also use single quote for strings
my_object = MyClass()
my_list = [1, ‘hi’, [True, 5]]
my_bool = my_list[2][0]
```

Functions

```
def my_math_function(x):
    return 2 * x + 1

def my_function(my_param):
    print('my_param is... ' + my_param)
    my_var = my_math_function(4)
    print(my_var)
```

Loops

```
for element in my_list:
    do_something(element)

while my_boolean_variable:
    do_something()
```

Conditionals

```
if boolean_expression:
   do_something()
elif other_boolean_expresion:
   do_something_else()
else:
   do_another_thing()
```

Demo

Follow along at
https://repl.it/@diegolopez/demoworkshop1

Data Pipeline

1. Gather raw data
2. Structure data
3. Explore & preprocess data
4. Report

Data Preprocessing

Motivation

- Real world data is often noisy and dirty
- Cleaning up the data is an important step in improving model performance
- Data visualization should be done prior to model selection and training

Preprocessing Tasks

- Data cleaning
- Data transformation
- Data reduction
  - Data discretization

Kinds of Data

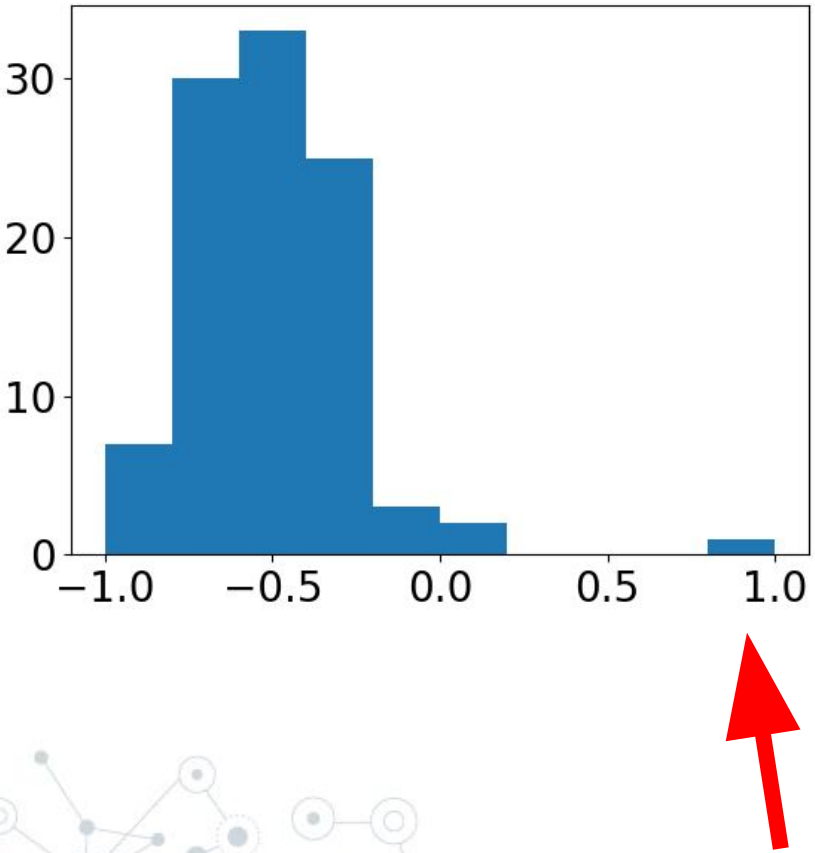
- Numbers
- Strings
  - Ignored in this workshop
- Categories
- Some mixture of the above

Dealing with Numbers

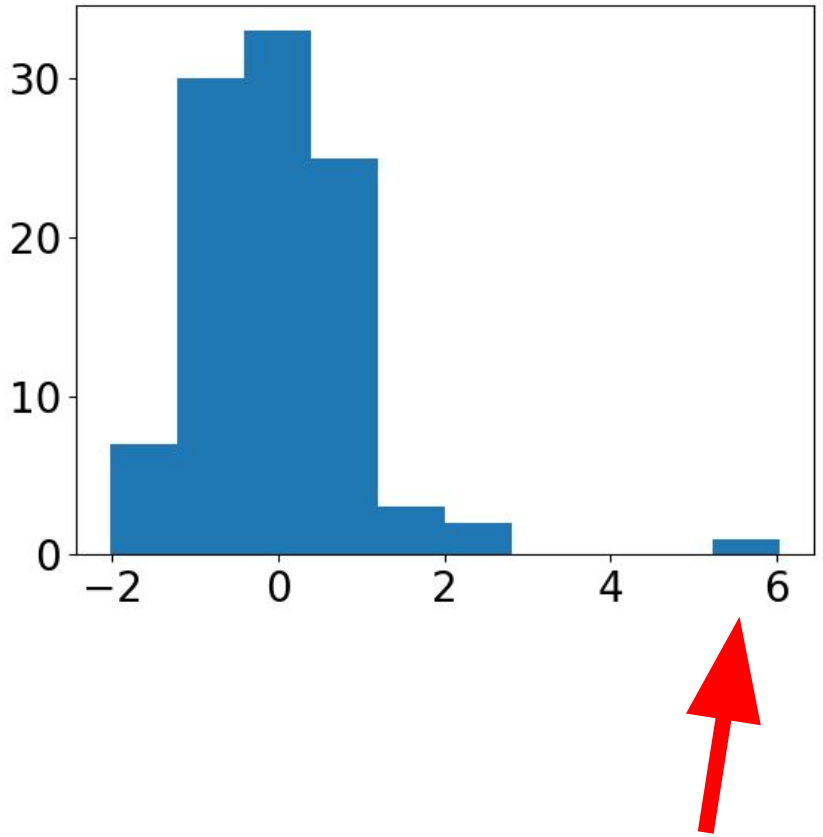
- Rescaling
  - Min-max normalization
    - Scale to $[0, 1]$ or $[-1, 1]$
  - Standardizing
  - Log scaling (for example, salary)
- Binerizing and bucketizing
  - “Discretizing” numbers into intervals
- NaNs
  - Replacing with mean
    - Could be done on a category by category basis
  - Predict missing values with a simple model
  - Disregarding data points with missing data

Scaling

Min-Max Scaling



Standard Scaling



Categorical Data

- Important to check for formatting differences
  - For instance, setting all words to lowercase
- One-hot encoding
  - Mapping each categorical value to a unit vector
  - If there are n categories, each we use n dimensional vectors, each unit vector corresponding to each category

Categorical Data

- Assume we a person can have one of 4 favourite colors: blue, red, yellow, and green.
- Then, the one-hot encoded vectors might look like this:

$$\vec{v}_{\text{blue}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \vec{v}_{\text{red}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \vec{v}_{\text{yellow}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \vec{v}_{\text{green}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Mixture

- Values with unit needs to be turned into a single number (and then rescaled accordingly)
- Dates can be turned into UNIX time (how much time has elapsed since Jan 1 1970?) and then rescaled
- Alternatively, dates can be bucketized into months

Adding Features

- If a feature is appropriately scaled, we can add features of higher degrees to add non-linearity
  - If x_1 and x_2 are features, we could add the feature ${x_1}^2$ or $x_1 x_2$
- Can be used for any model

Adding Features

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$

Advanced Techniques

- Text Data
  - Bag of words
  - n-gram
- Image Data
  - Data augmentation with image transformations
- Audio data
  - Fast fourier transform

The Dataset

- More than 15 thousand soccer players
- The columns include
  - Name
  - Nationality
  - Age
  - Height
  - Weight
  - Rating

Demo Time

Follow Along

tiny.cc/mais-f2019-w1-notebook

Thanks!

Check out our next workshop on Oct 9th:

Workshop 2: Deploying your ML app
facebook.com/events/401246070765344/

Also give us feedback on this workshop:
tiny.cc/MAIS-F2019-W1-feedback

Thanks!

Any questions?

You can find us at
https://mcgillai.com