INFO 6105
Data Science Engineering Methods and Tools

Lecture 6
Lab: Pima Indians Diabetes Data Set

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

R or Python?

Both languages are simple to learn and great for data science, but one might be better than the other depending upon your skills and your priorities.

- experience programming in other languages
- academia or industry
- data processing (pandas, data.table)
- data visualization (matplotlib, ggplot)
- machine learning (sciekit-learn, CARET)
- ...

Scikit-learn

Scikit-learn is a machine learning library for the Python programming language. It provides simple and efficient tools for

- Preprocessing: Feature extraction and normalization.
- Regression: Predicting a quantitative variable
- Classification: Predicting a qualitative variable
- Clustering: Automatic grouping of similar objects into sets.
- Model selection: Comparing, validating and choosing parameters and models.
- Dimensionality reduction: Reducing the number of random variables to consider.

How do I install scikit-learn?

- Option 1: Install scikit-learn library and dependencies (NumPy and SciPy)
- Option 2: Install Anaconda distribution of Python

Scikit-learn

Scikit-learn is a machine learning library for the Python programming language. It provides simple and efficient tools for

- Preprocessing: Feature extraction and normalization.
- Regression: Predicting a quantitative variable
- Classification: Predicting a qualitative variable
- Clustering: Automatic grouping of similar objects into sets.
- Model selection: Comparing, validating and choosing parameters and models.
- Dimensionality reduction: Reducing the number of random variables to consider.

How do I install scikit-learn?

- Option 1: Install scikit-learn library and dependencies (NumPy and SciPy)
- Option 2: Install Anaconda distribution of Python

scikit-learn

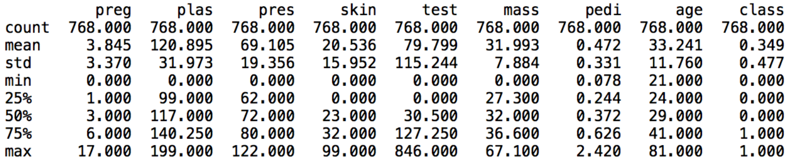
Requirements for working with data in scikit-learn

- Features and response should be separate objects
- Features and response should be numeric
- Features and response should be NumPy arrays
- Features and response should have specific shapes

Pima Indians Diabetes Data Set

- Sources: National Institute of Diabetes and Digestive and Kidney Diseases
- Objective: Predict based on diagnostic measurements whether a patient has diabetes
- Number of Instances: 768 (all females and at least 21 years old)
- Number of Features: 8 plus class
    1. Number of times pregnant
    2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
    3. Diastolic blood pressure (mm Hg)
    4. Triceps skin fold thickness (mm)
    5. 2-Hour serum insulin (mu U/ml)
    6. Body mass index (weight in kg/(height in m)^2)
    7. Diabetes pedigree function
    8. Age (years)
    9. Class variable (0 or 1)
- Download dataset: Machine Learning Repository

Descriptive Statistics



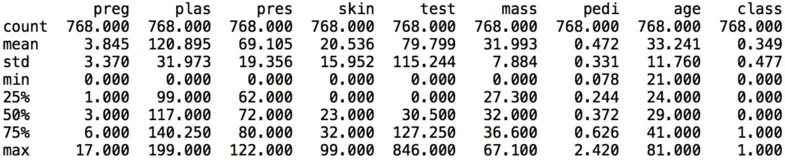
```
          preg     plas     pres     skin     test     mass     pedi      age    class
count  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000
mean     3.845  120.895   69.105   20.536   79.799   31.993    0.472   33.241    0.349
std      3.370   31.973   19.356   15.952  115.244    7.884    0.331   11.760    0.477
min      0.000    0.000    0.000    0.000    0.000    0.000    0.078   21.000    0.000
25%      1.000   99.000   62.000    0.000    0.000   27.300    0.244   24.000    0.000
50%      3.000  117.000   72.000   23.000   30.500   32.000    0.372   29.000    0.000
75%      6.000  140.250   80.000   32.000  127.250   36.600    0.626   41.000    1.000
max     17.000  199.000  122.000   99.000  846.000   67.100    2.420   81.000    1.000
```

Observations

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- Plasma glucose concentration: 5
- Diastolic blood pressure: 35
- Triceps skinfold thickness: 227
- 2-Hour serum insulin: 374
- Body mass index: 11

Descriptive Statistics



```
          preg     plas     pres     skin     test     mass     pedi      age    class
count  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000
mean     3.845  120.895   69.105   20.536   79.799   31.993    0.472   33.241    0.349
std      3.370   31.973   19.356   15.952  115.244    7.884    0.331   11.760    0.477
min      0.000    0.000    0.000    0.000    0.000    0.000    0.078   21.000    0.000
25%      1.000   99.000   62.000    0.000    0.000   27.300    0.244   24.000    0.000
50%      3.000  117.000   72.000   23.000   30.500   32.000    0.372   29.000    0.000
75%      6.000  140.250   80.000   32.000  127.250   36.600    0.626   41.000    1.000
max     17.000  199.000  122.000   99.000  846.000   67.100    2.420   81.000    1.000
```

Observations

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- Plasma glucose concentration: 5
- Diastolic blood pressure: 35
- Triceps skinfold thickness: 227
- 2-Hour serum insulin: 374
- Body mass index: 11

Descriptive Statistics

Observations (Cont.)

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- We can see that columns 1,2 and 5 have just a few zero values, whereas columns 3 and 4 show a lot more, nearly half of the rows.
- This highlights that different "missing value" strategies may be needed for different columns, e.g. to ensure that there are still a sufficient number of records left to train a predictive model.

Missing Values

Why is it important?

- Wrong conclusions
- Some statistical models fail if there is missing values
- Leading to poor performance

How to handle missing values?

- Remove Rows With Missing Values
- Impute Missing Values (e.g., Use a model to replace missing values.)
  - A constant value that has meaning within the domain, such as 0, distinct from all other values.
  - A value from another randomly selected record.
  - A mean, median or mode value for the column.
  - A value estimated by another predictive model.

Missing Values

Why is it important?

- Wrong conclusions
- Some statistical models fail if there is missing values
- Leading to poor performance

How to handle missing values?

- Remove Rows With Missing Values
- Impute Missing Values (e.g., Use a model to replace missing values.)
  - A constant value that has meaning within the domain, such as 0, distinct from all other values.
  - A value from another randomly selected record.
  - A mean, median or mode value for the column.
  - A value estimated by another predictive model.

Lab

Link: Pima Indians Diabetes Data in Python