Introduction to Data Science and AI

Tung Kieu

April 8, 2023

1. Requirements

- **Integrated Development Environment (IDE)**: PyCharm.
- **Version Control**: Git and TortoiseGit.
- **Compiler & Interpreter**: Python 3 (WinPython on Windows or Anaconda on Linux).
- **Additional Libraries**: Pandas, NumPy, SciPy, Matplotlib, Sklearn, (and PyTorch).
- **Data Sets**: Iris and MNIST

1.1. Download and install all the items in the requirements.
1.2. Check the installation is correct.

```
print("Hello, world.")
```

1.3. Load data set IRIS by using NumPy.
1.4. Print the IRIS data set to console.
1.5. Make 5% values in IRIS to `nan`.
1.6. Preproces missing data (i.e. `nan`) by using all the methods in the lecture.
1.7. For each preprocessing method, use a classification model (e.g., naive Bayes) and evaluate the accuracy.
1.8. Repeat step 1.6 and 1.7 with 10% `nan` values.
1.9. Repeat step 1.6 and 1.7 with 15% `nan` values.
1.10. Repeat step 1.6 and 1.7 with 20% `nan` values.
1.11. Use min-max-normalization on the data set and use a classification model (e.g., naive Bayes) and evaluate the accuracy.
1.12. Use z-normalization on the data set and use a classification model (e.g., naive Bayes) and evaluate the accuracy.