Independent Study

Machine Learning Tutorial

Satyaraja Dasara (sdasara@iu.edu)

System Prerequisites:

This tutorial requires a laptop/PC with an NVIDIA GPU.
The NVIDIA GPU should preferably have compute capability higher than 3.5.
Make sure that the GPU drivers are updated.
On Windows, they can be done in the following manner:
Device Manager → Display Adapters → Nvidia GPU → Right Click → Update Drivers → Select Automatically for Updated Driver Software

Installation of Python distribution:

We shall first install Anaconda.
Anaconda is a distribution which supports the creation of Python environments and various IDE's like Jupyter Notebook and Spyder.
The following steps help you install Anaconda:

1. Go to this link: https://www.anaconda.com/distribution/
2. Download the Python 3.7 64-Bit version for your OS.
3. Follow the steps here: https://docs.anaconda.com/anaconda/install/windows/

Creation and set-up of virtual environment:

After installation, we create a new virtual environment with required version of Python and packages/libraries.
The following steps tell you how to do it:

1. Open **Anaconda Powershell Prompt**
2. Run the following command to create a virtual environment with Python 3.6 and name as mltut:

   ```
   conda create --name mltut python=3.6
   ```

3. Run this command to activate this environment(This should be done every time you wish to use this environment):
   ```
   conda activate mltut
   ```
   If you want to deactivate this environment, run the following command:
   ```
   conda deactivate
   ```
4. Go to the directory containing **requirements.txt** with following command:
   ```
   cd <DIR-PATH>
   ```
   Eg.
   ```
    cd C:\Users\Satya\Desktop\GitHub\hospital-los-predictor
   ```
5. Now we have to install all the packages/libraries which we require for the tutorial. The package names and versions are all present in **requirements.txt**

   The packages/libraries that we would be needing are:
   - pandas
   - numpy
   - matplotlib
   - scikit-learn
   - tensorflow-gpu
6. Run the following command to install all these packages:
   ```
   conda install --yes --file requirements.txt
   ```
7. This step usually takes some time. In case you are unable to install some packages, run the following command instead:
   ```
   pip install -r requirements.txt
   ```
8. Conda install is a safer command than pip install. Pip should be used only when you are unable to install packages using conda.

Prerequisite knowledge for the Tutorial:

1. This tutorial requires basic knowledge of programming concepts.
2. Knowledge of Python with Pandas, Numpy, Matplotlib and Sklearn.
3. Basic idea of Machine Learning and Deep Learning would be helpful but not required.

Required Data:

1. task1.csv (derived from MIMIC-III database)
2. task2.csv (derived from UCI Diabetes 130-US hospitals for years 1999-2008 Data Set)

Tasks:

This tutorial will consist primarily of two different tasks on the two different datasets.
Task 1 is a regression task (estimate the value of a variable) to predict LOS (length of stay).
Task 2 is a classification task (predict the class of a variable) to predict readmission into hospital for diabetic patients.
Both the tasks will make use of sklearn library for machine learning algorithms and tensorflow Keras for deep learning.

Pseudo-code for ML:

X contains all the independent columns.
Y contains only target column.

First split the data into train and test datasets

```
X_train,X_test,Y_train,Y_test = split(X,Y)
```

Train the model using the training data

```
model.fit(X_train,Y_train)
```

Predict on the testing data

```
predictions = model.predict(X_test)
```

Calculate the metric i.e. MAE for regression/Accuracy for classification

```
metric(Y_test,predictions)
```