Machine learning in python

By the ISU CSE Club

What is machine learning

Machine learning: predictive analytics using large volumes of data by employing algorithms that iteratively learn from that data.

Uses: credit-card fraud detection, self-driving cars, optical character recognition, and online shopping recommendations, and much much more.

Why you should care: Machine learning (and data science in general) is becoming increasingly useful skills in our industry. As the amount of data we produce grows, the need to gather and utilize that data grows as well. Because of this, jobs requiring experience or knowledge about data science are becoming abundant

Machine Learning Process

4 steps:

- **Ingest** - Gather data (this step is often harder than it seems)
- **Process** - Prepare or clean the data for use in machine learning (make your data useable)
- **Predict** - Train and score a machine-learning model (the fun part)
- **Visualize** - Visualize the output from the model (look at the results from the fun part)

Tools we will use

Python- used extensively in the data science community for machine learning and statistical analysis due to availability of thousands of open-source libraries, such as NumPy, Pandas, Matplotlib, and scikit-learn, that can be used to explore, transform, analyze and visualize data.

Azure notebooks- cloud based platform for building and running Jupyter notebooks. Allows to write python code without having to install and manage a jupyter server.

Jupyter- environment based on IPython that facilitates interactive programming and data analysis using python (as well as other programming languages).

Sections in this Demo

Ingest- create an Azure account, a new Azure Jupyter Notebook, and import our dataset.

Process- use Pandas to understand our data then clean and prepare the data

Predict- use Scikit-learn to build a regression-based machine-learning model using a train-test split

Visualize- use Matplotlib to visualize the output of our ML model

Cleaning and preparing data

Also known as data wrangling, data scientists have to be careful what datasets are used to train these machine-learning models.

Typical ways of data wrangling: removing duplicate rows, removing rows or columns with missing values or algorithmically replacing the missing values, “normalizing” data, selecting specific columns.

Building machine-learning model

Scikit-learn: built in support for popular regression, classification, and clustering algorithms and works with other python libraries.

Train-test split: A train-test split is simply a division of your data into two parts (often approximately 80%/20% split). The larger portion of the split is used to train a machine learning model on while the smaller portion is used to cross-validate your model.

In this lab we will build a machine-learning model utilizing on-time arrival data for a major U.S. airline to create a model that can be used for predicting whether a flight is likely to arrive on time.

Visualize data

Matplotlib: popular plotting and charting library for python to visualize the result. Matplotlib, like all of the other tools used in this demo, has many features not covered in this demo. It is a good tool to use for not only visualizing your resulting ML model but also to analyze your data before creating a model to determine how to clean and process your data.

Why should you visualize: Machine learning is often used to distill very large amounts of data into a format that is easier for a human to understand or a computer to implement. While YOU may have an adept understanding of the data after working with it, you will need to present your findings to people with less technical experience and/or help implement software to utilize your findings. Visualization is a powerful tool for both of these.