CMSC 671: Principles of Artificial Intelligence Python for AI

Dr. Paula Matuszek

Paula.Matuszek@villanova.edu

Paula.Matuszek@gmail.com

(610) 647-9789

Why Python for AI?

- The traditional AI languages are Lisp and Prolog
- When performance is a serious issue commonly used languages include C++, sometimes Java, and more recently GPU-based languages
- So why Python?
  - Lightweight startup, interpreter, IDLE.
  - Object-oriented
  - Useful built-in data structures and operations for symbolic processing: dictionaries, lists, sets, strings
  - Strong numeric processing for statistical processing: matrix operations, etc.

Python for AI

- In addition to the built-in capabilities, Python gets much of its power from libraries that can be imported to add capabilities
- A couple of basic ones you will want (and probably have)
  - NumPy: math functions for forarrays and matrices
  - matplotlib: plotting library for Python and NumPy.
- Poole and Mackworth have a large set of Python tools which implement many basic AI functions. They are starting points, not polished code. http://artint.info/AIPython/ The PDF http://artint.info/AIPython/aipython.pdf gives a discussion of Python and of the tools.

More Specific AI Libraries

- As you get into more advanced topics, for projects or later in the course, you will reach some areas where you don't want to implement from scratch. There are a number of more specific libraries that let you concentrate on a broader problem.
- Some of the best-known are
  - Natural Language Toolkit (NLTK)
  - Scikit-learn
  - TensofFlow

NLTK

- What is NLTK?
  - Natural Language ToolKit
  - Set of modules for Python
  - Large number of tools for processing natural text
  - A largest of relevant data
  - Widely used for natural language recognition, speech processing, text mining, speech translation.
- Starting point is http://www.nltk.org/.

Some NLTK Modules

- NLTK is a toolkit for processing text
  - Text is treated as a list of words
- Modules include
  - Stemmers, Tokenizers, Parsers
  - Part of Speech and Named Entity Taggers
  - N-Grams, Frequency Distributions, other statistical
  - Classifying documents into predefined groups
  - Clustering documents into groups

NLTK Corpora

- NLTK also has a large set of corpora: existing text bodies that can readily be processed.
  - Brown: About 500 English documents, about a million words, compiled from a variety of sources. First general computer corpus available.
  - Reuters: about 800,000 news articles
  - Gutenberg: 18 works from 12 authors
  - Shakespeare: 8 of his plays
  - Current list at http://www.nltk.org/nltk_data/
- It also includes dictionaries, gazetteers, trained models.

Example

- Simple classifier, trying to classify names into male and female by the last letter.
- http://www.csc.villanova.edu/~matuszek/fall2013/Sep11Classify.py

Scikit-learn

- SciPy is "a Python-based ecosystem of open-source software for mathematics, science, and engineering." (https://www.scipy.org/)]
- NumPy and matplotlib are from here.
- Scikit-learn is a set of machine learning modules built on top of SciPy. (http://scikit-learn.org/stable/)

Scikit-learn modules

- Scikit-learn modules are grouped into six kinds:
  - Classification
  - Regression
  - Clustering
  - Dimensionality Reduction
  - Data Preparation
  - Model Selection

Classification

- Deciding what groip on class an object belongs to
- Algorithms:
  - Decision Trees
  - K Nearest Neighbor
    - Naive Bayes
    - Support Vector Machones
  - Examples: male vs female :-). Spam detection, loan applications, college admissions, image identification

Regression

- Predicting a continuous value (rather than a class) for an object
- Algorithms
  - Least squares linear regression
  - Logistic Regression
- Examples: which will this house sell for? What GPA can we expect this student to achieve? What temperature will it be tomorrow?

- Clustering: grouping similar objects without any a priori definition of groups
  - Algorithms:
    - K-Means
    - Agglomerative clustering
    - Hierarchical clustering
- Examples: what are the news topics in these articles? How many different things do I have images of? What are the kinds of calls to the help desk we are getting?

Dimensionality Reduction

- Reducing the total number of variables for each individual without losing explanatory power
- Algorithms
  - Principal Component Analysis
  - Factor Analysis
  - Singular Value Decomposition
- Examples. Reducing text vector lengths, eliminating some of the redundancy in natural language. Reducing image vectors, eliminating irrelevant variation from lighting.

The Rest

- Preprocessing. Data cleanup
  - Algorithms: feature extraction and normalization
  - Examples: Using weight and height in the same classification, turning a 10-point movie rating into "yes, no".
  - Model Selection. Checking usefulness of models, comparing different approaches and parameters.
  - Examples: Each of the above methods produces a mathematical model or formula which can then be applied to new data. And each has many parameters which can be tweaked. These tools help decide among the options and determine whether they are actually good.
- All of this is well documented at the scikit-learn site.

And others

- TensorFlow is Google's deep learning framework; it is written mostly in C++, but the API is Python based. It's relatively new, not as widely used as NLTK or Scikit-learn. But it's getting well-known (because Google). https://www.tensorflow.org/
- For any given AI problem it's worth checking GitHub or Googling to see what's available.
- Things written in Java or C++ may still have Python bindings available.
- Have fun!