Prolegomena unit:

- Course introduction (Monday)
- Basich machine learning terminology (**today**)
- Lab: Python libraries (Friday)
- (Start *The nature of data* on Wed, Jan 18)

Today:

- Review ML in context
- Fundamental vocabulary
- A classification example
- A regression example

There's a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.

Joel Grus, Data Science from Scratch, 2015, pg 1

he problem of searching for patterns in data is a fundamental one and has a long and successful history. For instance, the extensive astronomical observations of Tycho Brahe in the 16th century allowed Johannes Kepler to discover the empirical laws of planetary motion, which in turn provided a springboard for the development of classical mechanics. Similary, the discovery of regularities in atomic spectra played a key role in the development and verification of quantum physics in the early twentieth century. The field of pattern matching [used as a synonym for machine learning] is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.

Bishop, Pattern Recognition and Machine Learning, 2006, pg 1

Suppose we are studying ways to identify a tree's species by looking at its leaves. Associate each task below with the field of study that is the best fit for that task.

- Curating a large data set of images of leaves and making it usable for plant identification.
- Using a series of diagnostic questions to derive an application for identifying the species of a tree from attributes of a leaf. (For example, is the leaf simple of compound, does it have lobes, how thick is it..?)
- Finding the distribution of leaf length, thickness, and density with error margins.
- Using data to derive an application for identifying plant species from images of leaves.

- Eigenvalues and eigenvectors (Student)
- Vector and matrix multiplication (Student)
- Vector spaces (Student)
- Basic Python programming (Student)
- Python libraries for machine learning (CSCI 381)
- Derivatives (Student)
- Newton's method (Student)
- Partial derivatives (CSCI 381)
- Big-oh and big-theta (Student)
- Data structures for lists, maps, and sets (Student)
- Random variables (CSCI 381)
- Mean and variance (CSCI 381)
- Bayes's theorem (CSCI 381)

Machine learning main tasks:

- Regression, where the target type is $\mathbb{R}$
- Classification, where the target type is a finite set
  - Binary classification, where the target is $\{F, T\}$ (or $\{0, 1\}$ or $\{-1, 1\}$ ...)
- Density estimation, where the target type is $[0, 1]$.

Other machine learning tasks (see Goodfellow, *Deep Learning*, pg 96–100):

- Transcription, where the observations are unstructured and the targets are text.
- Machine translation, where the observations and targets are text.
- Anomaly detection, where the targets are indicators of whether the observation is atypical.
- Synthesis and sampling, where there are no observations in deployment, but rather the program produces new observations similar to those in training.
- Denoising, where the targets are corrected versions of the observations.

Coming up:

Learn Python

Take basic-terminology quiz (due classtime Friday)

Do Python warm-up assignment (due end-of-day Friday)