Lecture 1

1 Course introduction

Data science encapsulates multiple sub-disciplines related to the processing of data, the extraction of useful information from data, and the ability to make predictions using data. It includes:

- *Data mining:* Non-trivial extraction of useful, new, hidden, and/or implicit infor- mation from data.
- *Machine learning:* Field of study that gives computers the ability to learn without being explicitly programmed.
- *Big data:* Extremely large data sets that may be analyzed computationally to re- veal patterns, trends, and associa- tions, especially relating to human behavior and interactions.
- *Signal processing:* Processing, extracting, and transferring information contained in multitude different formats, broadly referred to as signals.

1.1 Supervised learning

Adapted from [1, Chapter 1.1].

Supervised learning deals with the following problem: suppose we are given data samples,

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y},$$

where $\mathcal{X}$ consists of the data measurements and $\mathcal{Y}$ are the corresponding labels. Some examples are:

- $\mathcal{X} =$ images
  $\mathcal{Y} =$ the dominant object in the image (e.g., cat, dog, pine tree, etc...)
- $\mathcal{X} =$ music samples
  $\mathcal{Y} =$ the musical genre

- $\mathcal{X}$ = molecules
  $\mathcal{Y}$ = the energy of the molecule
- $\mathcal{X}$ = scientific papers
  $\mathcal{Y}$ = the scientific field the paper belongs to (e.g., math, biology, physics, etc...)
- $\mathcal{X}$ = tweets
  $\mathcal{Y}$ = whether the tweet is positive or negative
- $\mathcal{X}$ = people's dogs
  $\mathcal{Y}$ = the species of dog (e.g.,)

Note $\mathcal{Y}$ can have two discrete classes (binary classification), more than two discrete classes (multiclass classification), or a continuum of values (regression). We have placed no assumptions on $\mathcal{X}$, other than it be a set.

In learning theory though, we need additional structure on $\mathcal{X} \times \mathcal{Y}$ so we can *generalize* from the samples $\{(x_i, y_i)\}_{i \leq n}$. This means, given a new data point $x \in \mathcal{X}$, we want to predict the corresponding label $y(x) = y \in \mathcal{Y}$. This requires defining a *similarity measure* between points $x, x' \in \mathcal{X}$. Defining a notion of similarity on $\mathcal{X}$ is a deep and multifaceted question at the heart of data science and machine learning.

1.2 Unsupervised learning

In unsupervised learning we remove the labels $\mathcal{Y}$. The goal of unsupervised learning algorithms is to find hidden patterns in the data $\mathcal{X}$. By its nature, this is a fuzzier goal than in supervised learning, in which there is a clear metric for success (the classification or regression error). Unsupervised learning can involve for example, clustering the data or performing dimension reduction. The latter entails finding low dimensional coordinates of the data that in a suitably defined sense lose little information content. The notion of similarity between x, x' often still plays a key role in these algorithms, as it guides the algorithm to place more importance on certain patterns over others.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.