Unsupervised Learning
Introduction
Unsupervised Learning

What can we learn from data when label information is not available?
Supervised learning framework

Data: \((x_1, y_1), (x_2, y_2), \ldots \in \mathcal{X} \times \mathcal{Y}\)

Assumption: there is a (relatively simple) function \(f^* : \mathcal{X} \rightarrow \mathcal{Y}\)

such that \(f^*(x_i) = y_i\) for most \(i\)

Learning task: given \(n\) examples from the data, find an approximation \(\hat{f} \approx f^*\)

Goal: \(\hat{f}\) gives mostly correct prediction on unseen examples
Unsupervised Learning

Data: \( \vec{x}_1, \vec{x}_2, \ldots \in \mathcal{X} \)

Assumption: there is an underlying structure in \( \mathcal{X} \)

Learning task: discover the structure given \( n \) examples from the data

Goal: come up with a summary of the data using the discovered structure

- **Partition the data into meaningful structures**
  - clustering
- **Find a representation that retains important information and suppresses irrelevant/noise information**
  - Representation, Embeddings, Dim. reduction
- **Understand and model how data is distributed**
  - Data analysis, density estimation
- **Processing techniques that aid in data analysis and prediction**
  - ad-hoc techniques

Some techniques do multiple tasks at once
A quick overview of the topics
Clustering

- Centroid based methods (k-centers, k-means, k-mediods,...)
- Graph based methods (spectral clustering)
- Hierarchical methods (Cluster trees, linkage based methods)
- Density based methods (DBSCAN, watershed methods)
- Bayesian methods (Mixture modelling, Dirichlet and Chinese Restaurant processes)
- Axiomatic frameworks (impossibility results)
Representations

- Metric Embeddings (metric spaces into $L_p$ spaces)
- Representations in Euclidean spaces (text and speech embeddings, vision)
- Representations in non-Euclidean spaces (hyperbolic embeddings)
- Dim. reduction in Euclidean spaces
  - linear methods (PCA, ICA, factor analysis, dictionary learning)
  - non-linear methods (LLE, IsoMap, t-SNE, autoencoders)
Data analysis and density estimation

- Parametric and nonparametric density estimation (classical techniques, VAEs GANs)
- Geometric data analysis (horseshoe effect, topological data analysis, etc.)
Ad-hoc techniques

- Organizing data for better prediction
  - Datastructures for nearest neighbors (Cover trees, LSH)
  - Datastructures for prediction (RPTrees)
This course: Goals

• To study in detail various methodologies applied in an **unsupervised learning** task

• Gain a deep **understanding** and working knowledge of the core theory behind the various approaches.
Prerequisites

Mathematical prerequisites

• Good understanding of: Prob and stats, Linear algebra, Calculus
• Basic understanding of: Analysis
• Nice to know: topology and diff. geom. (only for a few topics)

Computational prerequisites

• Basics of algorithms and datastructure design
• Ability to program in a high-level language.

Machine Learning prerequisites

• Good understanding of:
  Nearest neighbors, decision trees, SVMs, learning theory, regression, latent variable models, neural networks
Website:

http://www.cs.columbia.edu/~verma/classes/uml/

The team:

Instructor: Nakul Verma (me)

TA(s)

Students: you!

Evaluation:

• Homeworks (50%)
• Project (30%)
• Class participation (5%)
• Scribing and in class presentations (15%)
Homeworks (about 3 or 4 homeworks)

- No late homework
- **Must** type your homework (no handwritten homework)
- Must include your name and UNI
- Submit a pdf copy of the assignment via gradescope
- All homeworks will be done individually

- We encourage discussing the problems (piazza), but **please don’t copy**.

Project (can/should be done in a group of 3-4 students)

- Survey, implementation or some theory work on a specific topic in UL
- Details will be sent out soon
Class participation & Scribing

• Students should be prepared for class by reading the papers ahead of time
• Should actively participate in the class discussions

• Should present on one of the lecture topics covered in class
• (scribing) Should prepare a preliminary set of notes before the lecture, and update these notes with the detailed discussions that happen in class
Announcement!

• Visit the course website

• Review the basics (prerequisites)

• HW0 is out!

• Sign up on Piazza & Gradescope
Let’s get started!