Unsupervised learning in the age of Big DATA

Marina Meila
Department of Statistics
University of Washington
Supervised, Reinforcement, Unsupervised Learning

- We are witnessing an AI/ML revolution
  - this is led by Supervised and Reinforcement Learning
  - i.e. Prediction and Acting

- Unsupervised learning (clustering analysis, dimension reduction, explanatory models) in a much more primitive state of development
  - Everybody does them
  - Exploration, explanation, understanding
  - Is the next big challenge
Unsupervised learning is the next big challenge

Research in my group

- Unsupervised learning at scale
  - Clustering
  - Dimension reduction
  - Models for preferences

- Mathematics/theory/theorems/models
  - validation/checking/guarantees

- Algorithms and computation

- Geometry
  - Non-linear dimension reduction
  - Topological data analysis

- Combinatorics
  - Graphs, rankings
  - Clustering
Unsupervised learning is the next big challenge

Research in my group

- Unsupervised learning at scale
  - Clustering
  - Dimension reduction
  - Models for preferences

- Mathematics/theory/theorems/models
  - validation/checking/guarantees

- Algorithms and computation

- Geometry
  - Non-linear dimension reduction
  - Topological data analysis

- Combinatorics
  - Graphs, rankings
  - Clustering
Unsupervised learning is the next big challenge for the sciences.

Research in my group

- **Unsupervised learning at scale**
  - Clustering
  - Dimension reduction
  - Models for preferences
- Mathematics/theory/theorems/models
  - validation/checking/guarantees
- Algorithms and computation
- Geometry
  - Non-linear dimension reduction
  - Topological data analysis
- Combinatorics
  - Graphs, rankings
  - Clustering
Machine Learning/AI in the picture
(statistics, optimization, theoretical computer science)

Artificial intelligence

Speech recognition
Image captioning
Self driving cars
Translation
Playing chess

Hard sciences

Finance
Health
Power grid control
Robotics

Neuroscience
Biology
Chemistry
Astronomy
Physics

“What a human can do in about 1 second”
-- Andrew Ng cca 2019
**Machine Learning in the picture**

(Statistics, optimization, theoretical computer science)

<table>
<thead>
<tr>
<th>Artificial intelligence</th>
<th>Hard sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech recognition</td>
<td>Finance</td>
</tr>
<tr>
<td>Image captioning</td>
<td>Health</td>
</tr>
<tr>
<td>Self driving cars</td>
<td>Power grid control</td>
</tr>
<tr>
<td>Translation</td>
<td>Robotics</td>
</tr>
<tr>
<td>Playing chess</td>
<td>Neuroscience</td>
</tr>
<tr>
<td></td>
<td>Biology</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
</tr>
<tr>
<td></td>
<td>Astronomy</td>
</tr>
<tr>
<td></td>
<td>Physics</td>
</tr>
</tbody>
</table>

Correct?

Results EASY to validate

Results HARD to validate
Scientific discovery by machine learning and the mythical human “expert”

- Big data
  - Allows us to ask more detailed questions (e.g. “personalized medicine”)
  - Big data contains more complex patterns
  - Machine Learning discovers patterns fast
- Typically – validation by “domain experts”
- Often Hypotheses are cheap, experiments are expensive
Drowning in hypotheses...

Validation is the bottleneck

- Validation by visualization
- is qualitative not quantitative
- hard/impossible in dimension > 3
Drowning in hypotheses...

Validation is the bottleneck

- Validation by visualization
- is qualitative not quantitative
- hard/impossible in dimension > 3

- can’t be crowdsourced
Drowning in hypotheses...

Validation is the bottleneck
- Validation by visualization
- is qualitative not quantitative
- hard/impossible in dimension $\geq 3$
- can’t be crowdsourced
Manifold Learning
non-linear dimension reduction

- When?
  - Data in high dimensions
  - Data can be described by a small number of parameters
  - Large sample size necessary – for consistency
Manifold Learning
non-linear dimension reduction

■ When?
  ■ Data in high dimensions
  ■ Data can be described by a small number of parameters
  ■ Large sample size necessary – for consistency

■ Problems?
  ■ “Too expensive to perform for large data sets”
  ■ Results not comparable between algorithms, “good only for visualization”
Manifold Learning and Clustering for Millions of Points

https://www.github.com/megaman

megaman: Manifold Learning for Millions of Points

megaman is a scalable manifold learning package implemented in python. It has a front-end API designed to be familiar to actl learn but harnesses the C++ Fast Library for Approximate Nearest Neighbors (FLANN) and the Sparse Symmetric Positive Definite (SPDP) solver Locally Optimal Block Precondition Gradient (LOBPCG) method to scale manifold learning algorithms to large data sets. On a personal computer megaman can embed 1 million data points with hundreds of dimensions in 10 minutes. megaman is designed for researchers and as such caches intermediary stops and indices to allow for fast re-computation with new parameters.

Package documentation can be found at http://mmp2.github.io/megaman/

You can also find our arXiv paper at http://arxiv.org/abs/1603.02763

Examples

- Tutorial Notebook

Installation with Conda

The easiest way to install megaman and its dependencies is with conda, the cross-platform package manager for the scientific Python ecosystem.

James McQueen  Jake VanderPlas  Jerry Zhang  Grace Telford  Yu-chia Chen
Manifold Learning for Millions of Points

- English words and phrases taken from Google news (3,000,000 phrases originally represented in 300 dimensions by the Deep Neural Network word2vec [Mikolov et al].

- Main sample of galaxy spectra from the Sloan Digital Sky Survey (675,000 spectra originally in 3750 dimensions).
  preprocessed by Jake VanderPlas, figure by Grace Telford
Galaxy spectra from the Sloan Digital Sky Survey (675,000 spectra originally in 3750 dimensions).

The spectrum of a galaxy

Sloan Digital Sky Survey: where the spectra are from in the Universe
Distortions in Manifold Learning and how to remove them

- We estimate the distortion! (called push-forward Riemannian metric)

**Isomap** ML algorithm

**LTSA (Local Tangent Space Alignment)** ML algorithm
Exploring the configurations of small molecules

Configuration space of the Aspirin molecule (210,000 states x 21 atoms x 3 dim) after non-linear embedding with Diffusion Maps, colored by the torsion of the CH$_3$•C=O bond.

With Alexandre Tkatchenko and Stefan Chmiela.

“Polymorphic landscapes of molecular crystals” with A. Tkatchenko, R. DiStasio, A. Vasquez-Mayagoitia
Optimize neighborhood size [NIPS 2016]

Distances, angles, areas preserved

Choose independent e-vectors [NeurIPS 2019]

Riemannian relaxation [NIPS 2015]

Vector fields preserved

Estimate Riemannian metric

Coordinates with physical meaning

\[ Y_{i+1} = \text{grad}_T \phi_i (\xi_i) \]

\[ A_{i, \varphi} = \text{Proj}_T (\xi_i - \xi_{i'}) \]
Directed graph embedding
Manifold + vector field [NIPS 2011]

1-Laplacian estimation

Helmholtz-Hodge decomposition

Smoothed vector fields

Independent loops
(H_1 basis)

(In progress)
ManifoldLasso: coordinates with physical meaning
Model free guarantees for clustering

- Given a “good” clustering $C$ of a data set, prove that there is no other good clustering $C'$ too different from $C$. 

![Clustered data points diagram](image)
Model free guarantees for clustering

- Given a “good” clustering $C$ of a data set, prove that there is no other good clustering $C'$ too different from $C$. 

- Good and stable

- Bad and unstable

- Good and unstable
Model free guarantees for clustering

- **Framework:**

- Given a "good" clustering $C$ of a data set, prove that there is no other good clustering $C'$ too different from $C$. 
Clustering with data driven guarantees

\[ \text{CH}_3\text{Cl}+\text{Cl}^- \leftrightarrow \text{CH}_3\text{Cl}+\text{Cl}^- \]

MD simulation at T=900K
6 atoms x 3 dim

with Jim Pfaendtner and Chris Fu
Modeling Preferences

Preferential data is
- Discrete
- Many valued
- Non-Euclidean
- Has algebraic/combinatorial structure

Goal: do "statistics as usual" on large preference data
- E.g. what is the mean? Variance?
- Clustering? Regression? Bayesian inference?
- Estimate the structure of preferences
Statistics with rankings

- Modeling permutations by counting inversions
  - Flexible models, with interpretable parameters
  - Allow for efficient computation when consensus exists
  - Adapt to various types of missing data (e.g. top-k rankings, ratings, pairwise comparisons)

- Software [github.com/mmp2/dpmm-gmm](https://github.com/mmp2/dpmm-gmm)
  - C+matlab code performing Bayesian non-parametric clustering for ranked data

College programs $n = 533, N = 53737, t = 10$

DC116  DC114  DC111  DC148  DB512  DN021  LM054  WD048  LM020  LM050
WD028  DN008  TR071  DN012  DN052
FT491  FT353  FT471  FT541  FT402  FT404  TR004  FT351  FT110  FT352
Degree programs preference data: the clusters found

- Found more compact/homogeneous clusters than previous attempts
- Very large and very small clusters

- CS/Eng Dublin
- Architecture
- Home economics
Degree programs preference data: points vs. preferences

Students pay price of exam success as points jump

- Within each cluster, preferences do not depend on “grades/points” only

Masterclass students set new record for grades

Minister insists school subjects are not being ‘dumbed down’
The Structure of preferences

\[ N = 5000 \textbf{people ranked} \ n = 12 \textbf{types of sushi} \]

sake | ebi | ika | uni | tamago | kappa-maki | tekka-maki | anago | toro | maguro
ebi | kappa-maki | tamago | ika | toro | maguro | tekka-maki | anago | sake | uni
toro | ebi | maguro | ika | tekka-maki | uni | sake | anago | kappa-maki | tamago
tekka-maki | tamago | sake | ebi | ika | kappa-maki | maguro | toro | uni | anago
tamago | maguro | kappa-maki | ebi | sake | anago | uni | tekka-maki | toro | ika
uni | toro | ebi | anago | maguro | tekka-maki | ika | sake | kappa-maki | tamago
maguro | ika | toro | tekka-maki | ebi | uni | sake | tamago | anago | kappa-maki

- Preferences have \textbf{hierarchical structure}
- This was \textbf{estimated from data} (along with consensus and dispersions)
- Current work: \textbf{partial rankings}

sake | ebi,ika |
uni | toro,ebi,anago | maguro,tekka-maki |
The Structure of preferences

- Applied to peer review, surveys, social choice
- Upcoming Electoral Geometry and Gerrymandering group
Summary -- next challenges

- Finding explanations / descriptions
  - Unsupervised learning

- Validation
  - Of explanations
  - Of scientific hypotheses
  - Is much more costly than generating hypotheses (requires more data, new experiments, expert involvement)
Summary -- next challenges

- Finding explanations / descriptions
  - Unsupervised learning

- Validation
  - Of explanations
  - Of scientific hypotheses
  - Is much more costly than generating hypotheses (requires more data, new experiments, expert involvement)

- With BIG Data – machines must assist humans in validation process

- Mathematics must assist human intuition with non-Euclidean data

- Theory must assist computation (and conversely)
Ongoing and future projects

- Discovery in Materials Science and Molecular Chemistry, Active learning for material discovery (solar cells materials) (Alex)

- Discovering the structure of point clouds = geometric data analysis (James, Weicheng)
  - interpretable coordinates,
  - ML with vector fields,
  - finding the boundary of the data manifold,
  - manifolds with noise,
  - finding the loop basis and prime manifold decomposition

- Networks – which graph properties are stable/statistically significant?

- Modeling preferences and applications to peer review

- Clustering with data driven guarantees

...at the scale of the current data
What do my students do?

- Implement in python
- Apply to scientific data/problems (data analysis)
- Develop algorithms and methods
- Think geometrically
- Prove consistency or (sometimes) use other people’s proofs

What do they need to know?

- Be a reliable programmer
- 580s (some), multivariate analysis, non-parametric statistics
- Optimization/combinatorics/graph theory/CS algorithms/differential geometry – depending on the research topic
- Select ML areas (e.g. sparse regression) – go deeper as needed
- Willingness to learn new math or ML
Thank You!