CPSC 453 01 (FA21): Unsupervised Learning for Big Data

Jump to Today

CPSC/AMTH 453/553 (CBB/Gene 555)

This course covers machine-learning methods well-suited to tackling problems associated with analyzing high-dimensional, high-throughput noisy data including: nonlinear dimensionality reduction, kernels and data graphs, graph signal processing, clustering and coarse graining, information theoretic analysis, optimal transport, and neural network embeddings.

Students will be expected to complete three programming assignments throughout the semester. The assignments will be in the Python programming language. In addition to the programming assignments, there will be a final project and a final exam. Students can work in pairs. Students enrolled in scientific or engineering disciplines are strongly encouraged to select final projects related to their research interests. The course grade will be based on homeworks (30%), canvas quizzes (15%, 3% each x5), final project (30%), and final exam (25%). Please contact the instructor or teaching fellow in advance in order to request extensions. Extensions will not be granted less than 24 hours before the assignment is due. The course is suitable for upper-level undergraduates or graduate students in Computer Science, Genetics, Computational Biology & Bioinfomatics, or any science or engineering discipline. Students should have python programming experience and basic linear algebra (equivalent to MATH 222a or b or 225a or b, Linear Algebra). Students should consult with the instructor in advance on questions concerning background or prerequisites.

This year the course will take on a partially flipped format, with small pre-recorded lecture modules available online in the media library.

Instructor : Professor Smita Krishnaswamy **smita.krishnaswamy@yale.edu (mailto:smita.krishnaswamy@yale.edu)**

Class timings/location: 17 Hillhouse Room 101

Office: Tuesday 1pm (at AKW 104) or by appointment (email me) zoom.us/my/smitakrishnaswamy

TA: Sasha Safonova **sasha.safonova@yale.edu (mailto:smita.krishnaswamy@yale.edu)**

Mondays at 2 pm http://yale.zoom.us/my/sashasafonova

ULAs: Michal Gerasimiuk **michal.gerasimiuk@yale.edu (mailto:smita.krishnaswamy@yale.edu)** , Yasin Tarabar **yasin.tarabar@yale.edu (mailto:smita.krishnaswamy@yale.edu)**

ULA office hours: Yasin 3-5pm Fridays, Michal 3-5pm Thursdays AKW 104.

Online References: Mathematics of Data Science, online at:
https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf (https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf)

Neural Networks and Deep Learning, online at:

http://neuralnetworksanddeeplearning.com (http://neuralnetworksanddeeplearning.com/)

https://www.deeplearningbook.org/front_matter.pdf (https://www.deeplearningbook.org/front_matter.pdf)

piazza.com/yale/fall2021/cpsc453cbb555cpsc553gene555

Class jamboard:

Part 1: **https://jamboard.google.com/d/10Aonfo8DI7epG_FYvtqnv4zXbToGA7n0jMTREuKjExw/edit?usp=sharing (https://jamboard.google.com/d/10Aonfo8DI7epG_FYvtqnv4zXbToGA7n0jMTREuKjExw/edit?usp=sharing)**

Part 2: https://jamboard.google.com/d/1qXPJ1kz7NggyZkXSca7wf8MEQnhGbmF22NQxv6YZYcI/edit?usp=sharing

Lectures:

Lecture 1: Introduction to big data

https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=1&_r=3&sq=Brynjolfsson&st=cse&scp=1 (https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=1&_r=3&sq=Brynjolfsson&st=cse&scp=1)

https://www.nature.com/articles/s41591-019-0727-5 (https://www.nature.com/articles/s41591-019-0727-5)

Lecture 2: Linear Algebra Review

Google colab notebook: https://colab.research.google.com/drive/1NoYeySw6PljpBNesY5Xv1f5bUEj_AKFH?usp=sharing (https://colab.research.google.com/drive/1NoYeySw6PljpBNesY5Xv1f5bUEj_AKFH?usp=sharing)

http://mlwiki.org/index.php/Power_Iteration

Lecture 3: Covariance and PCA1

Google colab notebook: https://colab.research.google.com/drive/1skzPYiu8yy6JEYeRewJeQrH1Z8tuy4SG?usp=sharing (https://colab.research.google.com/drive/1skzPYiu8yy6JEYeRewJeQrH1Z8tuy4SG?usp=sharing)

Lecture 4: Covariance Geometry and PCA Derivation

Lecture 5: PCA, SVD and Low Rank approximation

Google colab notebook: https://colab.research.google.com/drive/1Z8e-M6o5MT7oPNp3LDX7-vPJ84vc-3gd?usp=sharing (https://colab.research.google.com/drive/1Z8e-M6o5MT7oPNp3LDX7-vPJ84vc-3gd?usp=sharing)

Lecture 6: MDS, distances and inner products

https://colab.research.google.com/drive/1skzPYiu8yy6JEYeRewJeQrH1Z8tuy4SG?usp=sharing (https://colab.research.google.com/drive/1skzPYiu8yy6JEYeRewJeQrH1Z8tuy4SG?usp=sharing)

Lecture 7: Kernel PCA

https://colab.research.google.com/drive/1PU1Db_9_agDotF5J-F3vTReEA5LMNaDt?usp=sharing (https://colab.research.google.com/drive/1PU1Db_9_agDotF5J-F3vTReEA5LMNaDt?usp=sharing)

Lecture 8: Diffusion Maps

https://www.sciencedirect.com/science/article/pii/S1063520306000546 (under reading) (https://www.sciencedirect.com/science/article/pii/S1063520306000546)

Lecture 9: Diffusion Maps and PHATE

https://www.nature.com/articles/s41587-019-0336-3 (https://www.nature.com/articles/s41587-019-0336-3)

https://colab.research.google.com/drive/14P3xy7O4WNghf9PXdVs8boT9i-x9CHEX?usp=sharing (https://colab.research.google.com/drive/14P3xy7O4WNghf9PXdVs8boT9i-x9CHEX?usp=sharing)

Lecture 10: PHATE and tSNE

http://www.cs.toronto.edu/~hinton/absps/tsne.pdf (http://www.cs.toronto.edu/~hinton/absps/tsne.pdf)

https://krishnaswamylab.github.io/visualization_comparison/

Lecture 11: Graph Laplacian and Graph Signal Processing

https://arxiv.org/abs/1211.0053 (https://arxiv.org/abs/1211.0053)

Lecture 12: Graph Filtering and Final project

https://pubmed.ncbi.nlm.nih.gov/29961576/ (https://pubmed.ncbi.nlm.nih.gov/29961576/)

Lecture 13: Graph Fourier and Wavelet Transforms

http://proceedings.mlr.press/v97/gao19e/gao19e.pdf (http://proceedings.mlr.press/v97/gao19e/gao19e.pdf)

https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34 (https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34)

https://mauromaggioni.duckdns.org/Papers/DiffusionWavelets.pdf (https://mauromaggioni.duckdns.org/Papers/DiffusionWavelets.pdf)

Lecture 14: Graph Dictionary Learning, Intro to Clustering

https://sites.fas.harvard.edu/~cs278/papers/ksvd.pdf (https://sites.fas.harvard.edu/~cs278/papers/ksvd.pdf)

https://arxiv.org/abs/1401.0887 (https://arxiv.org/abs/1401.0887)

Lecture 15: Kmeans and Spectral Clustering

https://towardsdatascience.com/spectral-clustering-aba2640c0d5b (https://towardsdatascience.com/spectral-clustering-aba2640c0d5b)

https://people.csail.mit.edu/dsontag/courses/ml14/notes/Luxburg07_tutorial_spectral_clustering.pdf (https://people.csail.mit.edu/dsontag/courses/ml14/notes/Luxburg07_tutorial_spectral_clustering.pdf)

https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf (https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf)

https://colab.research.google.com/github/KrishnaswamyLab/SingleCellWorkshop/blob/master/exercises/Clustering/n (https://colab.research.google.com/github/KrishnaswamyLab/SingleCellWorkshop/blob/master/exercises/Clustering/notebooks/01

Lecture 16: Hierarchical and Louvain

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 (https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68)

https://arxiv.org/abs/0803.0476

Lecture 17: Probability theory and kernel density estimation

https://towardsdatascience.com/histograms-vs-kdes-explained-ed62e7753f12 (https://towardsdatascience.com/histograms-vs-kdes-explained-ed62e7753f12)

https://scikit-learn.org/stable/modules/density.html (https://scikit-learn.org/stable/modules/density.html)

Lecture 18: Entropy, Mutual Information

Notebook from my workshop:

https://colab.research.google.com/drive/1_cYS8Lr8pt0HvAgdRNMcA-RsnHDui91B?usp=sharing

https://www.science.org/doi/10.1126/science.1250689#:~:text=The%20conditional%20density%20enables%20us,conc (https://www.science.org/doi/10.1126/science.1250689#:~:text=The%20conditional%20density%20enables%20us,concentrated%20

.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html)

Lecture 19: Comparing probability distributions

https://arxiv.org/pdf/1803.00567.pdf (https://arxiv.org/pdf/1803.00567.pdf) (Computational optimal transport book chapter 2, 4)

https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf (https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf) (Sinkhorn Algorithm)

Lecture 20: Intro to Neural Networks

https://towardsdatascience.com/how-to-code-a-simple-neural-network-in-pytorch-for-absolute-beginners-8f5209c50fdd (https://towardsdatascience.com/how-to-code-a-simple-neural-network-in-pytorch-for-absolute-beginners-8f5209c50fdd)

https://colab.research.google.com/github/KrishnaswamyLab/SingleCellWorkshop/blob/master/exercises/Deep_Learni
(https://colab.research.google.com/github/KrishnaswamyLab/SingleCellWorkshop/blob/master/exercises/Deep_Learning/notebook

Deep learning Book chapter 1, 2

Lecture 21: Stochastic gradient descent and backpropagation

Deep learning book chapter 1, 2

http://www.cs.toronto.edu/~hinton/absps/momentum.pdf (http://www.cs.toronto.edu/~hinton/absps/momentum.pdf)

Lecture 22: Autoencoders

https://www.cs.toronto.edu/~hinton/science.pdf (https://www.cs.toronto.edu/~hinton/science.pdf)

https://www.deeplearningbook.org/contents/autoencoders.html

Lecture 23: Variational Autoencoders and CNNs

https://arxiv.org/pdf/1606.05908.pdf (https://arxiv.org/pdf/1606.05908.pdf)

https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73#:~:text=variational%20autoencoders%20(VAEs)%20are%20autoencoders,order%20to%20ensure%20a%20bet

https://www.nature.com/articles/s41592-018-0229-2 (https://www.nature.com/articles/s41592-018-0229-2)

https://arxiv.org/abs/1505.04597

Lecture 24: GANs and variants

(GAN paper) https://arxiv.org/abs/1406.2661 (https://arxiv.org/abs/1406.2661)

(WGAN paper) https://arxiv.org/abs/1701.07875 (https://arxiv.org/abs/1701.07875)

(MAGAN paper) https://arxiv.org/abs/1803.00385

Lecture 25: Sequence models and wordvector embeddings

- Luong et al. Effective Approaches to Attention-based Neural Machine Translation 2015
- Hochreiter & Schmidhuber Long Sort Term Memory, 1997 (https://medium.com/datadriveninvestor/attention-in-rnns-321fbcd64f05)
- https://medium.com/datadriveninvestor/attention-in-rnns-321fbcd64f05 (https://medium.com/datadriveninvestor/attention-in-rnns-321fbcd64f05)
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)
- https://arxiv.org/abs/1301.3781 (https://arxiv.org/abs/1301.3781) (Word2vec paper)
- https://arxiv.org/abs/1706.03762 (https://arxiv.org/abs/1706.03762) (Transformer paper)

Lecture 26: Graph Neural Networks

http://web.stanford.edu/class/cs224w/ (http://web.stanford.edu/class/cs224w/)

https://www.youtube.com/watch?v=8owQBFAHw7E (https://www.youtube.com/watch?v=8owQBFAHw7E)

(https://www.youtube.com/watch?v=8owQBFAHw7E)

https://proceedings.mlr.press/v97/abu-el-haija19a/abu-el-haija19a.pdf

https://arxiv.org/abs/1710.10903

Course Summary:

Date	Details	Due
Sun Oct 10, 2021	Problem Set 1: Dimensionality Reduction via PCA and Diffusion Maps (https://yale.instructure.com/courses/68209/assignments/251970)	due by 11:59pm
Mon Nov 1, 2021	Quiz 2: Graph signal processing (https://yale.instructure.com/courses/68209/assignments/258276)	due by 11:59pm
Wed Nov 10, 2021	Problem Set 2: Graph Signal Processing and Clustering (https://yale.instructure.com/courses/68209/assignments/256991)	due by 11:59pm
Sun Nov 14, 2021	Project proposal (https://yale.instructure.com/courses/68209/assignments/256789)	due by 11:59pm
Sun Nov 21, 2021	Clustering and information theory quiz (https://yale.instructure.com/courses/68209/assignments/263034)	due by 11:59pm
Sun Dec 5, 2021	Problem Set 3: Feed-Forward Neural Networks, Autoencoders, Generative Models and Information Theory (https://yale.instructure.com/courses/68209/assignments/261922)	due by 11:59pm
Wed Dec 22, 2021	Final Exam (https://yale.instructure.com/courses/68209/assignments/266751)	due by 11:59pm
	Final project (https://yale.instructure.com/courses/68209/assignments/266007)	due by 11:59pm