DS-GA 1003 Machine Learning

# Overview

Rapid advances and deployment of machine learning based systems have dramatically changed how we view machine learning. We now frequently use stochastic gradient descent, or its variant, to train large-scale models with hundreds of millions, if not billions, of parameters on large datasets consisting of millions if not billions of training examples. We now train large-scale generative models from which we can readily and efficiently draw realistic samples, such as fluent natural language text as well as photorealistic images. Based on these generative models, trained on unlabelled examples, we can now train highly-accurate classifiers with only a few labelled examples. Behind all these wonders are simple yet foundational ideas rooted in mathematics, probability, statistics and computer science. In this course, we focus on these foundational concepts so that after taking this course, students will be able to understand modern, large-scale machine learning algorithms more systematically and rigorously.

# Target Audience

M.Sc. students and early-year PhD students in data science.

* Prerequisites: students should have taken the following courses.
  + DS-GA 1001 Introduction to Data Science
  + DS-GA 1002 Probability and Statistics for Data Science
  + Undergraduate-level courses on linear algebra, multivariate calculus, and probability

# General Information

* **Lectures:** 1h40m
  + Lectures will be in person, although they will be livestreamed via Zoom as well.
* **Lab sessions**: 50m
  + Lab sessions will be in person, although they will be livestreamed via Zoom as well.
* **Instructor**: [Kyunghyun Cho](http://www.kyunghyuncho.me/)
* **Assistant**: [Divyam Madaan](https://dmadaan.com/), [Bing Yan](https://bingyan.me/), [Michael Hu](https://michahu.github.io/) and [Lavender Jiang](https://lavenderjiang.github.io/)
* **Office Hours**
  + **Instructor**
    - Kyunghyun Cho: 1hr
  + **Assistant**
    - Divyam Madaan: 1hr
    - Bing Yan: 1hr
    - Michael Hu: 1hr
    - Lavender Jiang: 1hr
* **Grading**
  + Paper presentation 50%
  + Remote final exam 50%
* **Course Site**:
  + **Campuswire**: [REDACTED]
    - For discussion and announcements
    - Registration code: [REDACTED]
  + **Google Calendar**:
    - [REDACTED]
    - Add this to your calendar so that you get the up-to-date course schedule.
* **Lecture Note**:
  + <https://arxiv.org/abs/2505.03861>
  + This lecture note will be continuously updated throughout the semester.

# Schedule

Note that the schedule below is only a guideline. The content of each lecture will be decided as the course progresses.

| wk | Lecture (Tuesday) | Lab (Wednesday) |
| --- | --- | --- |
| 01/21 | - Logistics  - A roadmap of the course  - Energy functions: what ML is for us.  - Classification: perceptron and max-margin classifiers | - Setup: Github Repository.  - Options:  (a) Bring your own laptop  (b) Lightning studio  (All: DM) |
|
| 01/28 | Classification  - Prediction: softmax and cross-entropy  - Backpropagation | - Perceptron - Max-margin classifier - Logistic regression  (LJ) |
| 02/04 | Classification  - Stochastic gradient descent  Model selection and generalization  - Empirical risk vs. expected risk | - Multilayer perceptron  (LJ) |
|
| 02/11 | Model selection and generalization  - Bias-variance trade-off  - Uncertainty in the error rate  Building blocks of deep neural networks  - Convolutional networks  ~~- Recurrent networks~~  Model selection and generalization  - Hyperparameter tuning | Convolutional networks for image classification  (DM) |
| 02/18 | President’s Day | |
| 02/25 | Building blocks of deep neural networks - Normalization - Attention  Probabilistic machine learning (1)  - Probabilistic interpretation of an energy function  - Variational inference | Hyperparameter tuning with random search vs. Bayesian optimization  (DM) |
|
| 03/03 | Probabilistic machine learning (2)  - Mixture of Gaussians  - k-means clustering  - Probabilistic principal component analysis | Variational mixture of Gaussians vs. K-means  (BY) |
|
| 03/10 | Probabilistic machine learning (3)  - Variational autoencoders  - Variance of Monte Carlo and importance sampling  Brief discussion on causality (if time permits) | Variational autoencoders vs. PCA  (BY) |
|
| 03/18 | Undirected graphical models (1)  - Restricted Boltzmann machines  - MCMC Sampling  Review | Metropolis-Hastings for sampling from VAE with a partially-observed situation.  (MH) |
| 03/25 | Spring break | |
| 04/01 | Recap  Undirected graphical models (2)  - Energy-based generative adversarial networks  - Autoregressive models  (DM) | Autoregressive model inference (MH) |
| 04/08 | Reinforcement learning  - Policy gradient | Ensemble of deep neural networks  - Accuracy vs. calibration  (DM) |
|
| 04/15 | Ensemble methods  - Bagging - Bayesian machine learning  - Boosting  Meta-learning  - Learning to initialize the parameters from other datasets | Finetuning a language model using reinforcement learning  (MH) |
| 04/22 | Poster presentation 1 (18 teams)  (KC, LI, BY) | Poster presentation 2 (7 teams)  (LJ, BY) |
| 04/29 | Poster presentation 3 (18 teams)  (KC, LJ, BY) | Poster presentation 4 (7 teams)  (KC, LJ, BY) |
| 05/06 | PAC-Bayes generalization bound  Recap and Q&A | Reading Day |
| 05/06 -05/12 | Online Final Exam:   - Released on [REDACTED]  - Closed on [REDACTED] | |

# Paper presentation

Students will form teams of four to six at the beginning of the semester by random assignment. Each team will be randomly assigned to a group of papers from the Paper List below, and present the selected papers, or a subset, as a poster presentation. The assignment of groups and papers will happen after February 3 2025 (which is the Add/Drop deadline.)

Students will have 2-3 minutes to present their poster, followed by around 2 minutes of Q&A.

## Paper list

For some of the papers in the list below, use <https://library.nyu.edu/services/computing/off-campus/> to access the full text articles if they are behind the paywall.

1. Ronald Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. Machine Learning. 1992. [Link](https://link.springer.com/article/10.1007/BF00992696)
2. Chris Bishop. Mixture density networks. Technical Report. 1994. [Link](https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/bishop-ncrg-94-004.pdf)
3. Yann LeCun , Leon Bottou , Genevieve B. Orr , and Klaus-Robert Muller. Efficient BackProp. 1998. [Link](https://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf): Focus on Section 4.
4. Geoff Hinton. Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation. 2002. [Link](https://ieeexplore.ieee.org/abstract/document/6789337)
5. Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. 2005. [Link](https://www.jmlr.org/papers/volume6/lawrence05a/lawrence05a.pdf)
6. Salakhutdinov & Hinton. Semantic hashing. International Journal of Approximate Reasoning. 2009. [Link](https://www.sciencedirect.com/science/article/pii/S0888613X08001813)
7. Ilin & Raiko. Practical approaches to principal component analysis in the presence of missing values. JMLR. 2010. [Link](https://www.jmlr.org/papers/volume11/ilin10a/ilin10a.pdf)
8. Balakrishnan & Chopra. Two of a Kind or The Ratings Game? Adaptive Pairwise Preferences and Latent Factor Models. IEEE International Conference on Data Mining. 2010. [Link](https://ieeexplore-ieee-org.proxy.library.nyu.edu/document/5694029)
9. Bergstra & Hinton. Random search for hyper-parameter optimization. JMLR. 2012. [Link](https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf)
10. Pascanu, Mikolov & Bengio. On the difficulty of training recurrent neural networks. ICML. 2013. [Link](https://proceedings.mlr.press/v28/pascanu13.pdf)
11. Warde-Farley, Goodfellow, Courville and Bengio. An empirical analysis of dropout in piecewise linear networks. Technical Report. 2013. [Link](https://arxiv.org/abs/1312.6197)
12. Kingma & Welling. Auto-Encoding Variational Bayes. ICLR. 2014. [Link](https://arxiv.org/abs/1312.6114)
13. Ranganath, Gerrish and Blei. Black Box Variational Inference. ICML. 2014. [Link](https://proceedings.mlr.press/v33/ranganath14.pdf)
14. Finn, Abbeel and Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML. 2017. [Link](https://proceedings.mlr.press/v70/finn17a/finn17a.pdf)
15. Dinh, Pascanu, Bengio and Bengio. Sharp minima can generalize for deep nets. ICML. 2017. [Link](https://proceedings.mlr.press/v70/dinh17b)
16. Welleck, Yao, Gai, Mao, Zhang & Cho. Loss Functions for Multiset Prediction. NeurIPS. 2018. [Link](https://proceedings.neurips.cc/paper_files/paper/2018/hash/fb3f76858cb38e5b7fd113e0bc1c0721-Abstract.html)
17. Zhao, Mathieu & LeCun. Energy-based Generative Adversarial Network. ICLR. 2017. [Link](https://arxiv.org/abs/1609.03126)
18. Peters, Niculae & Martins. Sparse Sequence-to-Sequence Models. ACL. 2019. [Link](https://arxiv.org/abs/1905.05702)
19. Perez, Kiela & Cho. Rissanen data analysis: Examining dataset characteristics via description length. ICML. 2021. [Link](https://proceedings.mlr.press/v139/perez21a.html)
20. Reed, Zolna, Parisotto et al., A Generalist Agent. TMLR. 2022. [Link](https://openreview.net/forum?id=1ikK0kHjvj)
21. Hoffmann. Probabilistic latent semantic analysis. UAI. 1999. [Link](https://dl.acm.org/doi/10.5555/2073796.2073829)

## Extra Information

* A student in this course is expected to act professionally. Please follow the academic policies at NYU. You can find them at <https://bulletins.nyu.edu/graduate/arts-science/academic-policies>.
  + This course is part of NYU Center for Data Science, and we encourage you to read <[What Does It Mean To Be A Good Community Member At CDS?](https://docs.google.com/document/d/1hi_lxsXJC50kUh8BwCTGifFMl9t22U3kqh5vgOnZ5nI/edit?tab=t.0)>.
* If you are in need of medical and/or counseling support, please reach out to NYU’s Wellness Exchange ([wellness.exchange@nyu.edu](mailto:wellness.exchange@nyu.edu), 212-443-9999).
* Academic accommodations are available for students with disabilities. Please visit the [Moses Center for Accessibility and Inclusive Culture](https://www.nyu.edu/life/global-inclusion-and-diversity/centers-and-communities/accessibility.html) for further information.