AI for Genomics

Course description:

Sponsored by the city of Montréal, with support from Mila – Quebec Artificial Intelligence Institute and IVADO (Institut de valorisation des données), the AI in Genomics program is a 12-week training that will allow participants to get hands-on experience in working with machine learning. The program will help to prepare participants with expertise in genomics to develop a foundational knowledge of advanced machine learning methodologies so that they can develop a better understanding of where and how these techniques could be used with genomics data.

Dates

1/20/2020-4/13/2020

Instructors:

- Joseph Paul Cohen (Program Scientific Advisor)
- Tariq Daouda
- Paul Bertin
- Julie Hussin
- Ahmad Pesaranghader
- Sydney Swaine-Simon

System to Q&A: https://piazza.com/class/k4rmacqhp136ae
Student enroll Code: aigenomics101

Lecture 1: Onboarding and Introduction to neural networks (Tariq Daouda, January 24th @ 15 h -18 h)

Participants should get a basic understanding of neural networks and deep learning as well as enough practical knowledge to start building neural networks.

- Datasets
- Classification
- KNN
- Regression
- Evaluation: Accuracy (train, test, validation)
- Basics of Backprop (momentum?)

- Fully connected layers
- Non-linearities (Relu, tanh, sigmoid)
- Conv (1D, 2D)
- pyTorch introduction (Colab)
- Practical: pyTorch feed forward: Fully connected & Conv

Example slides: link

Lecture 2: Representation learning and backprop (Joseph Paul Cohen, January 31st @ 15 h -18 h)

Location- John Molson School of Business - S2.445 - Classroom

Deep learning overview, representation learning methods in detail (sammons map, t-sne), the backprop algorithm in detail, and regularization and its impact on optimization.

- (30min) What is deep learning overview (Slides)
  - Define supervised and self-supervised prob perspective
  - How to approach problems (use sklearn)
  - Examples of go-to methods: logistic regression, decision tree etc (use sklearn)
- (45min) Backprop in more detail (Slides)
  - Work through an example of manually performing the algorithm
  - Backpropagation (visualizing the chain rule)
  - Intuition for applying gradient updates for arbitrary functions
- break
- (1hr) Representation learning (Slides)
  - Non-linear dim reduction
  - word2vec
  - Sammons map (tutorial code)
  - t-SNE
  - Regularization

Lecture 3: Challenges facing ML in genomics (Paul Bertin, February 7th @ 15 h -18 h)

Challenges facing machine learning and deep learning techniques when applied to genomics: biases (acquisition and population structure), high dimensionality, and interpretability/explanation of models.

- why deep learning works (learnt features etc) When to use it or not? (20min)
  - If you have hand crafted features just use them.
  - need to add prior knowledge in the pipeline

- interpretability
  - (saliency map, (integrated gradients) etc, show some code example? Have the notebook open) (30min)
  - machine learning is not a causal discovery tool, only correlations!
- Fat data (diet networks), limited number of example (SNPs and gene expression) (10min)
- dataset biases
  - (dropout in single cell, different depths of reads depending on the experiments, different libraries, different acquisition machines)
  - population structure (IID assumption, give intuitive examples related to genomics) (30min)
- Go over the details of the paper paper so that people have one example in mind : find a good paper working with gene expression! (the paper doing the gene graph conv? Use it for the practical??)
- Practical:
  - Implement saliency map and ~~integrated gradients~~. Highlight the limitations of those techniques for gene expression data (how noisy are the feature importances, try to replicate over several trials etc)
  - Put the attendees in a situation where population biases prevent any meaningful learning, and how to account for those biases (use TCGA with gender as confounder and cancer type as target, use meta benchmark)

Lecture 4: Deep Learning Models in Genomics (Ahmad Pesaranghader and Julie Hussin, February 14th @ 15 h -18 h)

In the first part of this lecture, we introduce the different DL architectures used in population and functional genomics. In the second part of this lecture, we then introduce generative models and explore how they can be beneficial in the context of genomics, mainly for the augmentation of the training data.

1. Deep learning in population genetics and multi-omics (1h)

- Introduction to population and functional genomics
- Simulations in population genetics.
- Convolutional Neural Networks (CNNs) for population genetics inference
- Motif-based approaches in functional genomics
- DeepSEA (link) and state-of-the art models in functional genomics.

2. Advanced deep learning models for genomics (1h30)

- Variational AutoEncoders (VAEs)

- Generative Adversarial Networks (GANs)
- Limitations of vanilla GANs and vanilla VAEs
- GANs and VAEs in Genomics
- Discussion of interesting applications in the field mainly with respect to different omics data-types (current state-of-the-art and guideline for future work)

3. Tutorial (30 mins)

- Quick Implementation of vanilla VAE/GAN in PyTorch (Google Colab)
- GANs from the Paper: Generating and designing DNA with deep generative models (https://arxiv.org/abs/1712.06148)

Lecture 5: Ethics
(Sydney Swaine-Simon, February 21st @ 15 h -18 h)

In this lecture we will discuss the ethics associated with Genomics data and developing machine learning algorithms.

More details will be announced soon

Office hours:

Office Hours #1: : March 6th @ 15 h -18 h
Office Hours #2: : March 13th @ 15 h -18 h
Office Hours #3: : March 20th @ 15 h -18 h Location: Centre for Teaching and Learning – FB 620
Office Hours #4: : March 27th @ 15 h -18 h
Office Hours #5: : April 3rd @ 15 h -18 h

Tutorials

- Focus on gene expression data
- Load data and explore the dataset
- Build a simple logistic regression model to predict the tissue
    - Regularisation concept
    - Produce diagnostic plot and assess overfitting
- Build a more complex model
- Provide a github repo which makes a nice “playground”: extend the TCGA-benchmark project with graph loaders. People will be able to download the data and graphs easily with academictorrents

Some papers and additional resources :

- https://github.com/gokceneraslan/awesome-deepbio
- https://github.com/hussius/deeplearning-biology
- Libbrecht MW et al. Machine learning applications in genetics and genomics, Nat.Rev.Genetic 2015
- Jiang P et al. Big data mining yields novel insights on cancer, Nat Genet. 2015
- Deep learning: new computational modelling techniques for genomics (2019) : https://www.nature.com/articles/s41576-019-0122-6
- A primer on deep learning in genomics : https://www.nature.com/articles/s41588-018-0295-5

Functional genomics papers:

- DeepSEA: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4768299/
- DeFine: https://www.ncbi.nlm.nih.gov/pubmed/29617928
- DanQ: https://www.ncbi.nlm.nih.gov/pubmed/27084946
- DeeperBind: https://arxiv.org/abs/1611.05777
- SPEID: https://link.springer.com/article/10.1007/s40484-019-0154-0

Population genetics papers:

- https://www.ncbi.nlm.nih.gov/pubmed/29331490
- https://www.ncbi.nlm.nih.gov/pubmed/30517664
- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004845
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2927-x

Torrente, Aurora, et al. “Identification of Cancer Related Genes Using a Comprehensive Map of Human Gene Expression.” *PLOS ONE*, edited by Paolo Provero, vol. 11, no. 6, Public Library of Science, June 2016, p. e0157484, doi:10.1371/journal.pone.0157484.

Ching, Travers, et al. “Opportunities And Obstacles For Deep Learning In Biology And Medicine.” *Journal of The Royal Society Interface*, Cold Spring Harbor Laboratory, Jan. 2018, doi:10.1101/142760.

https://canvas.stanford.edu/courses/51037