CSCI 1850: Deep Learning in Genomics (Spring 2021) **(Due: April 1, 2021 11:59 PM)**

Homework Assignment #3

Instructor: Ritambhara Singh
HTA: Daniel Ben-Isvy
UTA: Giselle Garcia
Email: cs1850tas@lists.brown.edu

About the assignment: This assignment has been designed to help you think critically about the topics we covered in the class, encourage you to look at the literature, and get familiar with applying deep learning frameworks on single-cell datasets.

For the conceptual questions, we are looking for answers that are **maximum** 3 lines per point. So if a question is worth 2 points try answering it in maximum 6 lines. When writing answers, try to first list the main idea addressing the question and then expand on it. You may refer to the papers in the Reference section while answering these questions.

For the programming assignment, we will run your code to check if it gives the correct output. If the code does not run successfully, we will assign partial scores to the correct logic behind the implementation.

Attempting bonus questions or tasks is encouraged but not required.

Total: 40 points + 10 bonus points

1 Conceptual Questions

[30 points + 4 bonus points]

Question 1: (3 points)

(a) [2 points] What are the challenges associated with data imputation?
(b) [1 point] Give two examples of missing data in the health/biology domain.

Question 2: (3 points)

How does the Avocado paper demonstrate that the latent representation it learns captures biologically relevant information?

Question 3: (10 points)

Compare and contrast the PREDICTD [1] and Avocado [2] approaches for the following:
(a) [2 points] Architecture
(b) [2 points] Functions (Think: are they linear operations or not?)
(c) [2 points] Combining Latent Factors (Think: do they sum them or not?)
(d) [2 points] Training (Think: number of parameters)
(e) [2 points] Performance

Question 4: (8 points)

[Refer [3]]
(a) [2 points] How does single-cell RNA-Seq (scRNA-Seq) information differ from regular (or bulk) RNA-Seq? What information could you obtain from single cell data that you could not obtain from sequencing a whole culture of cells?

(b) [4 points] List and discuss four challenges of working with single-cell genomics datasets.
(c) [2 points] Why are classical imputation methods with defined missing values not suitable for single-cell data?

Question 5: (6 points)

Compare and contrast the DeepCPG [4] and DCA [5] approaches for the following:
(a) [2 points] Architecture
(b) [2 points] Task
(c) [2 points] Application and Performance

Question 6 (bonus): (4 points)

(a) [1 point] What could be the potential reason that DCA [5] using Mean Squared Error (MSE) does not provide good de-noising results for the scRNA-seq data?
(b) [3 points] How do the authors solve this issue?

2 Programming Assignment

[10 points + 6 bonus points]

Background: Single-cell RNA-Seq (scRNA-Seq) data is incredibly noisy and sparse, with only a fraction of transcribed RNAs captured in the sequencing due to the difficulty in amplifying such a small signal from a single cell. Specifically, there are many genes expressed in a cell at any given time, and two entirely different scRNA-Seq readings could come from cells with identical expression profiles.

In this assignment, you will explore utilizing autoencoders to identify clusters of cells from the counts data. These clusters could represent varying expression profiles in the same tissue sample, such as the unique profiles that you might see in different phases of the cell cycle or different cell types.

Dataset: In the face of these many challenges, simulation is important for being able to conduct repeatable experiments with an actual ground truth to test the efficacy of various methods. The particular package used to generate this scRNA-Seq count data is known as Splatter[6], and it serves this purpose of recreating single-cell data for evaluation purposes. The data provided will be in the form of normalized log counts across 1000 genes for 5000 cells.

Task [10 points]:
Given the skeleton code:

1. **[4 points]** Train an Autoencoder (composed of fully-connected layers) to learn a low-level gene expression representation of the cells from the scRNA-Seq counts.
   (a) Use latent embedding (embedding learned by the bottleneck layer) sizes of 5, 10, 50, and 100. This is the size of the output of the encoder, and the input to the decoder. Also, follow the implementation notes in the *assignment.py* comments.
   (b) Report and compare the reconstruction Mean Squared Error (MSE) for each latent embedding size. How does the size of the latent space affect the reconstruction MSE?
2. **[6 points]** Analyze the learned encodings. We have provided code to produce PCA & t-SNE plots of the raw scRNA-Seq counts.
   (a) Adapt the provided code to make PCA & t-SNE plots of the autoencoder reconstructions.
   (b) Compare and report the plots of the original data with the plots of the reconstructions. How do the PCA & t-SNE plots of the reconstructions compare with those of the original data?
   (c) Take the trained encoders of the previous autoencoders and make PCA & t-SNE plots of the latent embedding vectors. Remember to make and submit plots for all 4 autoencoders (latent embedding sizes - 5, 10, 50, and 100). How does the size of the latent space affect the latent vector plots?
   (d) Compare and report the plots of the original data and the plots of the latent vectors. How do the various embedding sizes change the quality of clustering?

Bonus task [6 points]: Convert your autoencoder to a denoising autoencoder. (Reconstruction loss is calculated using the original input without the added noise.)

1. Add Gaussian noise to the input data and train the denoising autoencoder to remove the added artifical noise.

2. Add negative binomial distribution (zero-inflated negative binomial) noise to the input data and train the denoising autoencoder to remove the added artificial noise.

3. For each of the above, compare and report the PCA plots of the original data and the plots of the latent vectors. Is the de-noising autoencoder able to remove the artificial noise in your data?

References

[1] Timothy J Durham, Maxwell W Libbrecht, J Jeffry Howbert, Jeff Bilmes, and William Stafford Noble. Predictd parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature communications*, 9(1):1–15, 2018.

[2] Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *BioRxiv*, page 364976, 2019.

[3] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[4] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1):67, 2017.

[5] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

[6] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Course policies:

Collaboration Policy: Discussion of material with your classmates is both permitted and encouraged. However, showing, copying or other sharing of answers to written questions and actual code **for this assignment is forbidden**. This includes publishing projects publicly on Github or any other public platform. In addition, reusing code or pre-trained models from another student or a public platform **for this assignment is forbidden**.

Missed assignments (including late assignments): You can get a 3-day extension for at most 3 deadlines without penalty. Excluding the scenario mentioned above, 20% of the total points will be deducted for late submissions and missed submissions won't be assigned any score.