CSCI 1850: Deep Learning in Genomics
Spring 2021

http://cs.brown.edu/courses/csci1850

Mar 11, 2021
Thursday

Instructor: Ritambhara Singh
Format: Online (Synchronous)
Time: TTh 10:30-11:50 AM
Section III: Imputation of genomic signals

• LSTMs (revisit)

• Predicting methylation states in single cells (contd.) : DeepCpG

• Autoencoders

• Class Activity

• Denoising single-cell data using autoencoders: DCA
Recurrent Neural Networks (RNNs)

Useful for datasets with sequential dependency

$h_t = f(h_{t-1}, x_t)$

$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t)$
Long Short Term Memory Networks (LSTMs)

Unable to capture long-range dependencies

Cell state

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Long Short Term Memory Networks (LSTMs)

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Long Short Term Memory Networks (LSTMs)

She gave the book to **him** and **he**.

\[ f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f) \]

1: “completely keep this”

0: “completely get rid of this”

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Long Short Term Memory Networks (LSTMs)

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$LSTM$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
She gave the book to **him** and **he**...
Long Short Term Memory Networks (LSTMs)

She gave the book to him and he decided..

\[ x_{t-1} \quad x_t \quad x_{t+1} \]

- LSTM Cell state
- Output gate
- \( o_t = \sigma (W_o [h_{t-1}, x_t] + h_o) \)
- \( h_t = o_t \cdot \text{tanh} (C_t) \)

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Gated Recurrent Units (GRUs)

**GRU**

Cell state + Hidden state

Update gate = Input + Forget gate

**LSTM**

Image courtesy: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Questions?
DNA methylation

Change on DNA sequence and not histone tails

Known to be related to gene inactivation

Image Courtesy: DNA methylation and memory formation, Day and Sweatt 2010
DeepCpG: Input

**1: CpG site present**

**0: CpG site absent**

**?: Missing data**

CpG methylation states

25 nearest **known** CpG sites in both directions + distance [Dim = 100]

DNA sequence

One-hot encoding of 101 length sequence around the CpG site
DeepCpG: Model

Can we distinguish between “0” and “?”

Can we impute using this model?

Multi-task learning

Only joint module re-trained
Since the proportion of methylated versus unmethylated CpG sites can be unbalanced … we used the area under the receiver operating characteristics curve (AUC) to quantify the prediction performance of different models.”

18 mouse embryonic stem cells (mESCs)

<table>
<thead>
<tr>
<th></th>
<th>Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1, 3, 5, 7, 9, 11</td>
</tr>
<tr>
<td>Validation</td>
<td>13, 14, 15, 16, 17, 18, 19</td>
</tr>
<tr>
<td>Test</td>
<td>2, 4, 6, 8, 10, 12</td>
</tr>
</tbody>
</table>

Why the box plots?
“Since the proportion of methylated versus unmethylated CpG sites can be unbalanced … we used the area under the receiver operating characteristics curve (AUC) to quantify the prediction performance of different models.”

Could you modify the evaluation strategy to measure imputation performance?
Questions?
Autoencoders

We were learning a reduced embedding in Avocado too, what is different here?

Why the bottleneck?

\[
\phi : \mathcal{X} \rightarrow \mathcal{F} \\
\psi : \mathcal{F} \rightarrow \mathcal{X} \\
\phi, \psi = \arg\min_{\phi, \psi} \| X - (\psi \circ \phi) X \|^2
\]

Image courtesy: https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368
Denoising Autoencoders

What can this model be used for?

Discriminative/Generative modeling?

Image courtesy: https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368
Denoising Autoencoders is robust to noise

Input images from fashion MNIST.

Input images with salt and pepper noise.

Output from denoising network

Image courtesy: https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368
Questions?
Think:
- Given a single-cell count matrix (Fig (A)), how would you design a denoising autoencoder to remove noise for each cell? What would be the input and output? Size of the first layer MLP layer? [6 mins]

- What are Fig (B) examples of - multi-class classification or multi-task (also known as multi-label) classification? What activation function will you use for each and why? [4 mins]

Pair
Share: https://docs.google.com/document/d/1axNDTkuSB3jjOL4YxGehP8EpQg6OgYXp7rSbRCaIVHk/edit?usp=sharing
Questions?
Single-cell RNA-seq denoising using a deep count autoencoder

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller & Fabian J. Theis

Nature Communications 10, Article number: 390 (2019) | Cite this article
Noise in the single-cell data

- Low RNA capture rate leads to failure of detection of an expressed gene
- “False” zero count observation: defined as dropout event
- True zero counts represent the lack of expression of a gene in a specific cell-type.
- Not all zeros in scRNA-seq data can be considered missing values.

“Due to the non-trivial distinction between true and false zero counts, classical imputation methods with defined missing values may not be suitable for scRNA-seq data.”
Denoising

Manifold = Underlying Structure
What’s the simplest reconstruction loss that comes to mind?
DCA: Output

Replace MSE loss with ZINB loss

Represents the noise model
What is ZINB?

Zero inflated negative binomial distribution
What is ZINB? How do we model count data?

- Poisson distribution
- Mean = Variance

Remember Poisson regression?

Y can take multiple discrete read count values

Image courtesy: https://www.researchgate.net/figure/Poisson-distributions-for-different-values-of-lambda-Y-is-a-count-and-lambda-is-the_fig1_255717571
What is ZINB? How do we model count data?

- Poisson distribution: Mean = Variance
- RNA-Seq data can have variable variance due to noise (artifacts, unknown causes)
- Mean != Variance

Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50)

Courtesy: https://en.wikipedia.org/wiki/Variance#/media/File:Comparison_standard_deviations.svg
What is ZINB? How do we model count data?

- **Poisson distribution**: Mean = Variance
- **Negative Binomial**: Mean ≠ Variance

RNA-Seq data can have variable variance due to noise (artifacts, unknown causes).

Image courtesy: https://www.vosesoftware.com/riskwiki/NegativeBinomial.php
What is ZINB?

Zero inflated negative binomial distribution

Why zero inflated version for single-cell?

Image courtesy: https://stats.stackexchange.com/questions/264528/is-this-zero-inflated-negative-binomial-distribution
ZINB \( (x; \pi, \mu, \theta) = \pi \delta_0 (x) + (1 - \pi) \text{NB} (x; \mu, \theta) \)

\( \hat{\Pi}, \hat{M}, \hat{\Theta} = \arg\min_{\Pi, M, \Theta} \text{NLL}_{\text{ZINB}} (x; \Pi, M, \Theta) \)

DCA: Summary

Denoised output

Represents the noise model
DCA: Results

Simulations

- Without dropout
- With dropout
- Denoised (ZINB)
- Denoised (MSE)
Upcoming

Course website: [http://cs.brown.edu/courses/csci1850](http://cs.brown.edu/courses/csci1850)

Section III: Imputation of genomic signals

- Homework 3 will be released **next Tuesday (March 16)**

- Course project presentations **next week (March 16 and March 18)**
  - Only use histone marks for mid-term (inclusion of sequences is for Finals)
  - 7 min presentations w/ all members presenting + 3-5 mins of Q/A
  - Add our Github IDs to your private repos (6/10)
  - Replace “TBD” to team names
  - Submit your model results to the Kaggle site
Wrap up

What was the clearest point today?

What was the muddiest point today?

https://forms.gle/4B9iUxic2bpNDGM6A