Biomedical Data Science 2021 Homework 2

Due: May 16th (Sunday) 11:59pm EST

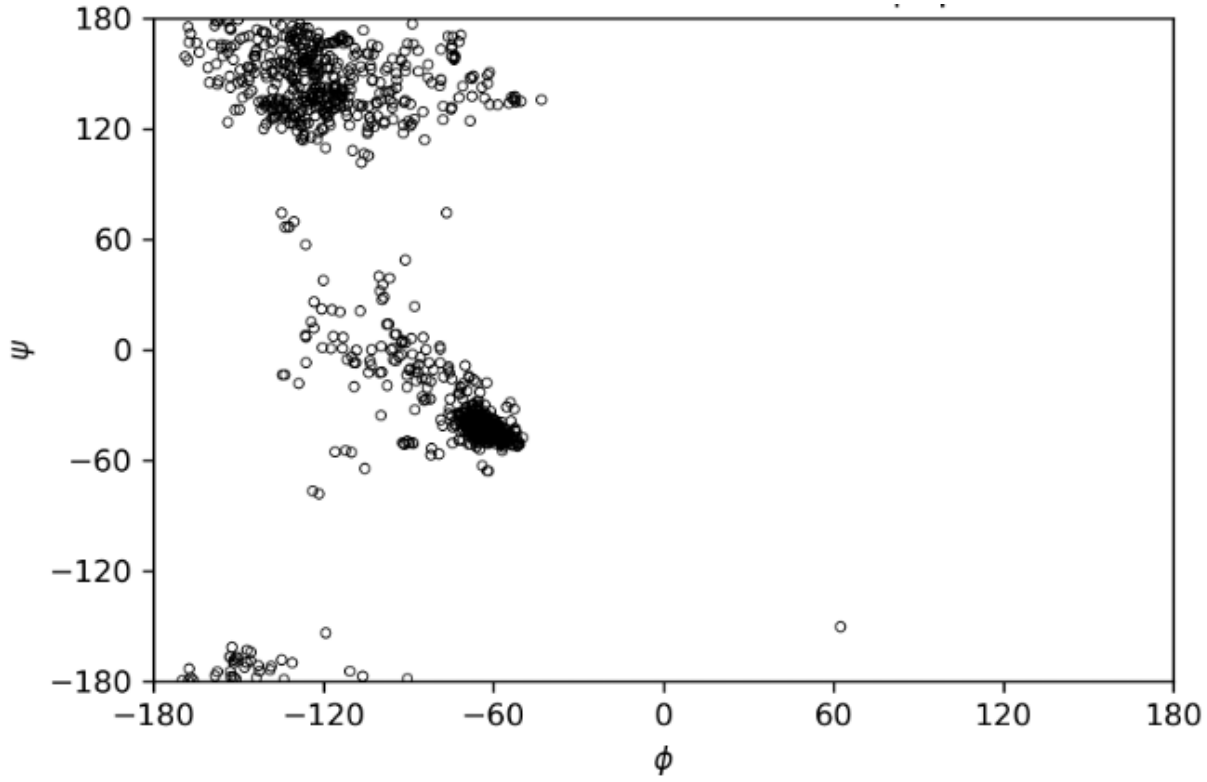
Submission should be done in Canvas

CBB & CPSC & S&DS (programming)

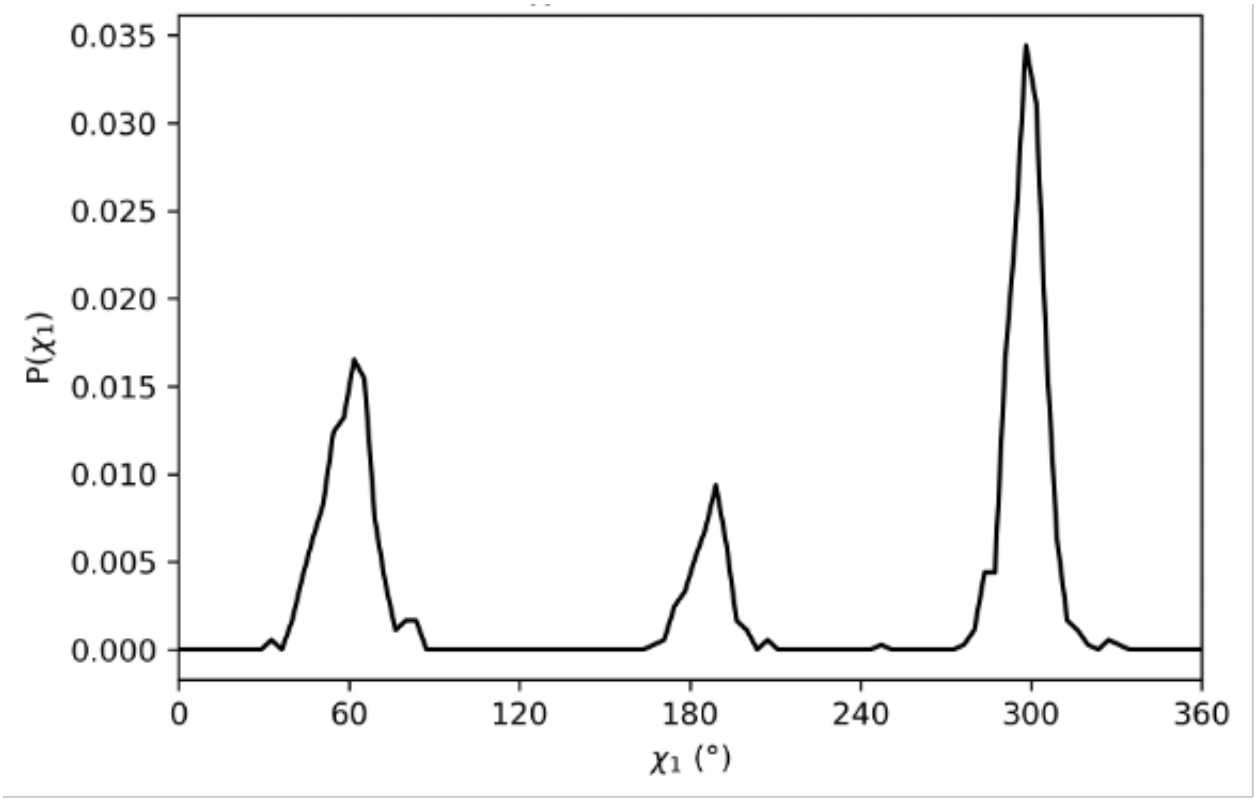
Submit a zip file including all the python/R codes for question 1 and 2, and the solution to question 3. All supplementary files could be find here or in Canvas files.

- Similar to homework 1, please indicate how to run your code in your code comments.
- You may submit the code in jupyter notebook format with the plot generated.
- If you make used of any online resource please cite the source in the code comments. You may use some small utility functions directly, but notice that directly copying large chunks of codes (even with variable name replacement) are not allowed and will be considered as plagrism.

1. (35pt) Ramachandran plots allow us to investigate the sterically allowed and disallowed backbone dihedral angle combinations ϕ and ψ in proteins. Using the file `core_THR_residues.txt` provided, produce a Ramachandran plot for threonine residues. The file `core_THR_residues.txt` contains 1000 threonine dipeptides taken from a database of high-resolution protein crystal structures. The Cα, caroboxyl carbon, and oxygen atoms on the prior amino acid are labelled pCa, pC, and pO. The N, Cα and H atoms on the subsequent amino acid are labeled: nN, nCa and nH. Using this file, calculate ϕ and ψ for each residue and produce a Ramachandran plot similar to that shown in below. See the lecture notes for definitions of ϕ and ψ.



2. (35pt) In the lecture notes, we not only discussed backbone dihedral angles ϕ and ψ, but we also discussed sidechain dihedral angles. As the side chains have different numbers of atoms, they can have different numbers of sidechain dihedral angles. In the case of threonine, there is only one sidechain dihedral angle χ1. Generate the observed side chain dihedral angle distribution from `core_THR_residues.txt` discussed in question 1. The observed distribution should be similar to that shown in below. See the lecture notes for the definition of χ1 in threonine.



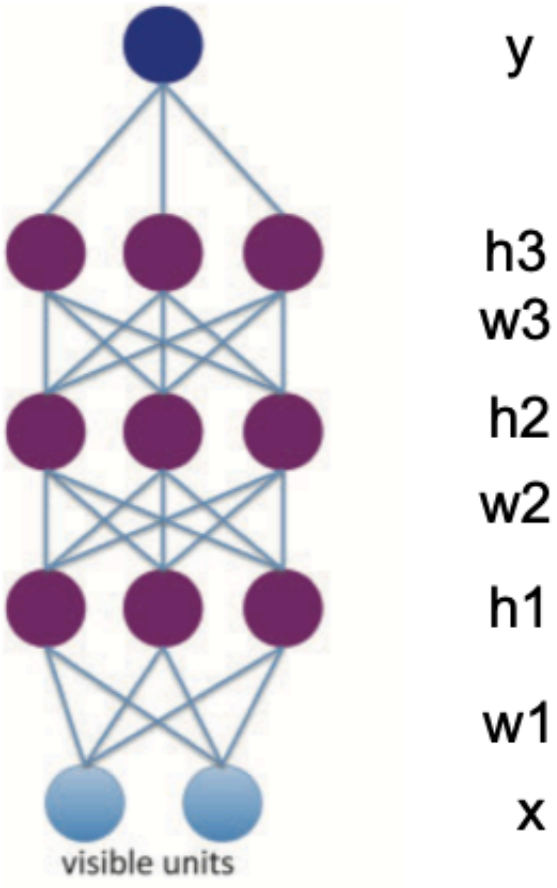
3. (30pt) Write down the cross-entropy loss function over Softmax output units, derive the gradient of the loss function with respect to the logit z in the Softmax, and explain why the cross-entropy loss is better than a squared error loss in this case.

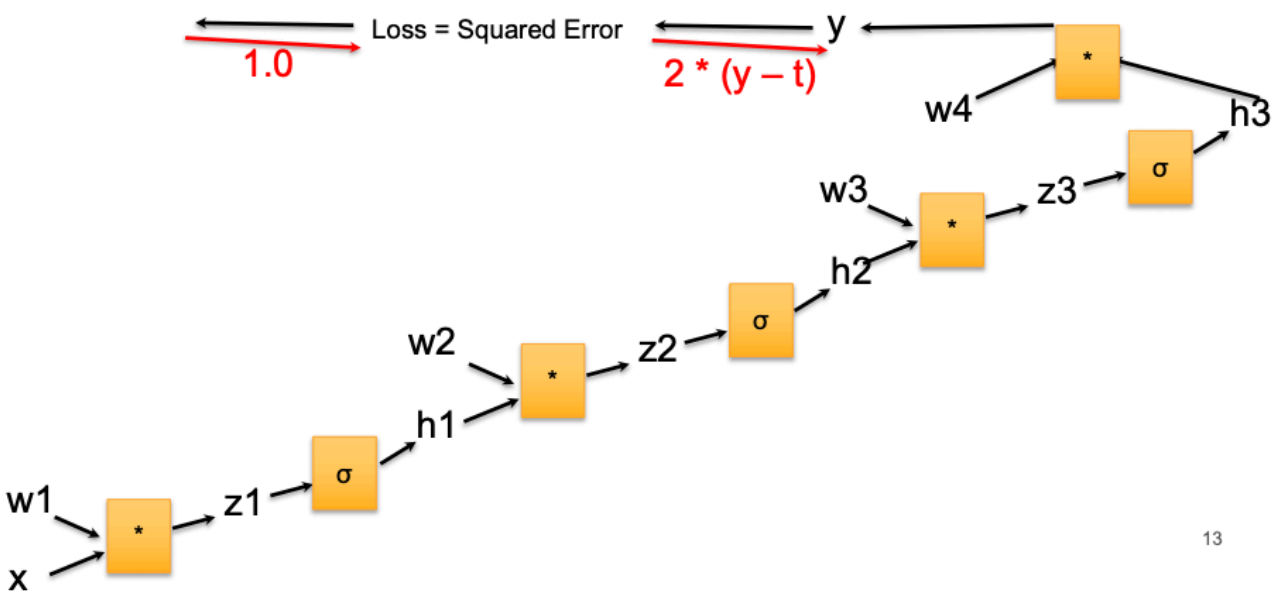
MBB&MCDB (non-programming)

Submit a single file answering the following questions.

1. (35pt) Derive the expressions for th x-, y-, and z-components of the force $\overrightarrow{F_j}$ on atom j=i+1 from the previous atom i and successive atom k=i+2 using the bond angle potential, $V_{ba} = \frac{k_\theta}{2}(\theta_{ijk} - \theta_0)^2$, where k_θ is the constant bond stiffness, $\theta_{ijk} = \cos^{-1}(\frac{\overrightarrow{r_{ij}} \cdot \overrightarrow{r_{kj}}}{r_{ij} r_{kj}})$ is the bond angle between bonded atoms i, j, and k, $\overrightarrow{r_{ij}} = \overrightarrow{r_i} - \overrightarrow{r_j}$, and θ_0 is the preferred bond angle. Note that $\overrightarrow{F_j} = \frac{-dV_{ba}}{dx_j}\hat{x}\frac{-dV_{ba}}{dy_j}\hat{y}\frac{-dV_{ba}}{dz_j}\hat{z}$.

2. (35pt) With the computation graph below, derive the gradient of the network with a linear output unit and a squared error loss based on backpropagation (with respect to the weight w1).





3. Read the following paper and write a short summary:

Grønbech, Christopher Heje, et al. "scVAE: Variational auto-encoders for single-cell gene expression data." *Bioinformatics* 36.16 (2020): 4415-4422.

In your summary, please try to answer these questions:

- What do the authors want to achieve?
- What is the major advantage of using variational autoencoders compared to other methods (esp. traditional autoencoders)?
- How is the input data represented (i.e. what is provided to the model)?
- How do the authors design the likelihood function and what is the intuition behind it?
- What experiments do the authors perform to show the effectiveness of the model?