Final Exam for I529, Spring 2017

Due by May 2nd (Tuesday), 11:59PM.

Instruction: This is a take-home exam. You should complete the exam independently (**no collaboration with your friends**). If you get your ideas from a web page or other resources, you need to provide proper citations. Push your answers to your IU github repository under a folder called final. **No late submission will be accepted**.

1. (10 points) Use your own words to describe what's stochastic gradient descent approach and how it is different from gradient descent approach used in ANN.
2. (20 points total) Given the following DNA sequences, assuming they are collected independently: ATCATC, CTATAG, GCATCG, ATCAGT, and AATCCG.
   a) Derive the 1st order Markov chain model for these sequences, and show the model in a matrix. (10 pts)
   b) For a new sequence ATCG, what's its probability given the Markov chain model? (5 pts)
   c) Compared to a random DNA model, is this sequence (ATCG) likely to be generated by the Markov chain model you just learned? Justify your answer (5 pts)

3. (40 points) Eukaryotic cells have compartments, and proteins are located in different compartments. Many factors play a role in deciding the destination of a protein. Build a predictor for protein location prediction using the tools that you have learned in this semester. For this question, you will use the Yeast data available at https://archive.ics.uci.edu/ml/datasets/Yeast as the training dataset. Make sure that you perform proper assessment of the performance of your predictor. Describe/report briefly and precisely the following: a) the package you use, b) the ML approach, c) how you do the cross-validation, d) how you evaluate the performance, e) the commands you use, and f) the performance of your predictor. Please also discuss briefly what else you would like to try (or how differently you would do) to build the predictor, if you were given more time to work on this problem.

4. (30 pts) Differential DNA composition and gene contents are often used for detecting alien DNAs (including phages, plasmids and other mobile genetic elements) that are integrated into bacterial genomes. For this problem, you are asked to devise an approach for predicting if a given bacterial genome contains any integrated phage(s) or plasmid (s), and if so where the alien DNAs are located within the bacterial genome using DNA composition difference. You may also discuss how you will incorporate the gene contents (e.g., phage genomes may contain genes encoding for viral structural proteins) for the prediction. You don't need to implement codes for this problem, but please give detailed description of your approach (what model you will use, how you will train your model and how you will do the testing?).