CSCI 5525: Machine Learning (Fall' 18)

Homework 4, Due 12/17/18

1. **(30 points)** This problem considers optimization methods for deep learning based on a mini-batch $\{(\mathbf{x}_1, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ of size m in each iteration. Let θ denote the parameter vector of the deep learning model.

   (a) (20 points) Clearly describe the Adagrad algorithm for iteratively computing the parameter vector θ. What limitations of the basic stochastic gradient descent approach does the Adagrad algorithm address and how? What are the limitations of the Adagrad approach for deep learning?

   (b) (10 points) Clearly explain why gradient based methods for learning parameters in a deep learning model may face the problem of exploding or vanishing gradients. Do you expect to see such problems more in the deeper layers closer to the input or shallower layers closer to the output? Briefly justify your answer.

2. **(40 Extra Credit points)** Describe an architecture, objective function and optimization procedure for a deep learning sytem that can solve one of the following problems. Be as specific as possible, and full credit goes for a model with a) equations b) an explanation for why it will work c) Tensorflow or PyTorch code for the **architecture**.

   i) Given a database of text queries and corresponding structured database queries (e.g. SQL), learn a function that translates text to SQL.

   ii) Given a dataset $D = \{X, y\}$ with X an MxN feature matrix of M instances and y a matrix of target vectors, assume you are also given $I = \{Z, z\}$, a binary matrix indicating a **non-random** pattern of missing feature information in X and y. Design a system that learns a probabilistic function relating X to y while also learning the missing data values. Assume the fraction of missing data is substantial.

   iii) Given fashionmnist https://www.kaggle.com/zalando-research/fashionmnist, *and* assume a **set of time series of fashion choices from a set of users,** design a network that can both classify the items AND predict user's future choices.

treat missing data equivalent to dropouts