CSE5312/CSEG312 Data Mining
Jihoon Yang
Spring 2021

Exam 2 (Take-home)
10:30∼11:45 a.m., June 10, 2021

Instructions

1. Please write your name in the space provided.
2. The test should contain 5 numbered pages. Make sure you have all 5 pages before you proceed.
3. Please consult the proctor if you have difficulty in *understanding* any of the problems.
4. Please be brief and precise in your answers. Write your solutions in the space provided.
5. Please show all the major steps in calculation to get partial credit where appropriate.
6. You have 75 minutes to complete the test. Good luck!

Name: **Student No.:**

Problem	Score
1	/20
2	/5
3	/10
4	/12
5	/7
6	/16
Total	/70

1. ($10 \times 2 = 20$ points) Multiple-choice/TF Questions (No explanations required):

   (a) Which of the following is/are true about bagging trees?

   i. In bagging trees, individual trees are independent of each other.

   ii. Bagging is a method for improving the performance by aggregating the result of weak learners.

   A) i   B) ii   C) i and ii   D) None of these

   (b) Which of the following is/are true about boosting trees?

   i. In boosting trees, individual trees are independent of each other.

   ii. Boosting is a method for improving the performance by aggregating the result of weak learners.

   A) i   B) ii   C) i and ii   D) None of these

   (c) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

   i. Both methods can be used for classification tasks.

   ii. Random Forest can be used for classification whereas Gradient Boosting can be used for regression tasks.

   iii. Random Forest can be used for regression whereas Gradient Boosting can be used for classification tasks.

   iv. Both methods can be used for regression tasks.

   A) i   B) ii   C) iii   D) iv   E) i and iv

   (d) In Random Forest you generate a set of trees and then aggregate the results of these trees. Which of the following is/are true about an individual tree in Random Forest?

   i. An individual tree is built on a subset of the features

   ii. An individual tree is built on all the features

   iii. An individual tree is built on a subset of the observations

   iv. An individual tree is built on full set of observations

   A) i and iii   B) i and iv   C) ii and iii   D) ii and iv

   (e) Which of the following is/are true about Gradient Boosting Trees?

   i. In each statge, introduce a new tree to compensate the shortcomings of existing model.

   ii. We can use gradient descent method to minimize the loss function.

   A) i   B) ii   C) i and ii   D) None of these

   (f) Bagging is suitable for high variance, low bias models.
   A) True   B) False

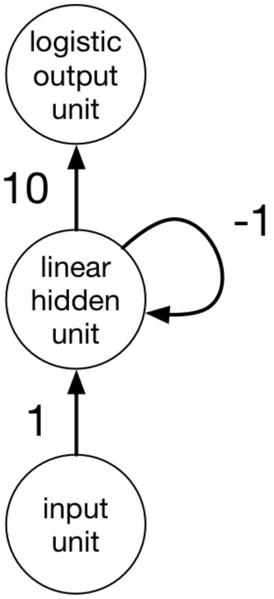
   (g) Let's say we have m number of estimators (trees) in a boosted tree. Now, how many intermediate trees will work on modified version (or weighted) of data set?
   A) 1   B) m-1   C) m D) Can't say   E) None of these

   (h) Boosted decision trees perform better than Logistic Regression on imbalanced class problems (e.g. anomaly detection problem).
   A) True, because they give more weight for lesser weighted class in successive rounds.
   B) False, because boosted trees are based on Decision Trees, which will try to overfit the data.

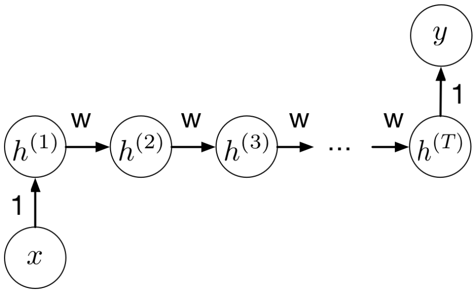
   (i) Provided $n < N$ and $m < M$. A Bagged Decision Tree with a dataset of N rows and M columns uses ___ rows and ___ columns for training an individual intermediate tree.
   A) N, M B) N, m C) n, M D) n, m

   (j) Prediction of individual trees of bagged decision trees has lower correlation in comparison to individual trees of random forest.
   A) True   B) False

2. (5 points) Clearly explain what the following network computes *at the final time step.* Assume that all biases are 0, and the inputs are integers and the length of the input sequence is even.



3. ($2 \times 5 = 10$ points) Given the following RNN with scalar input and output at the first and the last time step, with a shifted sigmoid function: $\phi(z) = \sigma(z) - 0.5$

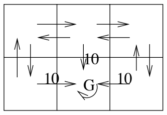


(a) Show the derivative $\overline{h_t}$ as a function of $\overline{h_{t+1}}$ for $t < T$, w, σ' (derivative of sigmoid function), and z_t (input to the activation function at time t).

(b) Assume $x = 0$ and thus $h_t = 0$ for all t. Compute the value α such that if $w < \alpha$, the gradient vanishes, while if $w > \alpha$, the gradient explodes. You may use the fact that $\sigma'(0) = 1/4$.

4. ($4 \times 3 = 12$ points) Count the total number of weights and connections in the first and second convolution layers in AlexNet.

5. (7 points) Consider the deterministic grid world shown below with the absorbing goal-state G in addition to five other states numbered 1 through 5. Here the immediate rewards are 10 for the labeled transitions and 0 for all unlabeled transitions.



1	2	3
4	G	5

Consider applying the Q learning algorithm to this grid world, assuming the table of $\hat{Q}$ values is initialized to zero. Assume the agent begins in the bottom right grid square and then travels counterclockwise around the perimeter of the grid until it reaches the absorbing goal state, completing the first training episode. Describe which $\hat{Q}$ values are modified as a result of this episode, and give their revised values. Answer the question again assuming the agent now performs a second identical episode. Assume $\gamma = 0.8$.

6. (16 points) Short-answer questions: give a succinct but clear answer.

   (a) Show and explain the rationale behind ELBO (Evidence Lower BOund) in VAE. (3 pts.)

(b) Show and explain the rationale behind the objective function of GAN. (3 pts.)

(c) Delineate the differences between REINFORCE (MC), REINFORCE with Baseline, and Actor-Critic algorithms. (3 pts.)

(d) Depict the architecture of the Transformer, and explain the role of each component. (7 pts.)