CS247: Advanced Data Mining (LEC80, Winter 2024)

Final Exam

Exam Date: **March 16, 2023**

Instructions

- The problem set will be available to you from 10:00AM to 12:00PM (noon) Pacific Time.
- The time limitation is **1 hour and 50 minutes**, plus 10 minutes for scanning and uploading your answers. Please note that submissions past 12:00PM will not be considered under any circumstances.
- The exam is online, open-book, and open-note.
- The midterm exam consists of two parts:
  - **Multiple-Choice Problems:** Please answer these directly on BruinLearn.
  - **Open-Answer Problems:** For these questions, upload your responses as **individual JPEG or PDF files**. Both typed and scanned handwritten answers are acceptable. If you do not have access to a scanner, you may also take a clear picture using your mobile phone. Please ensure the images are legible and **do allocate additional time** for this process if you choose to handwrite your answers.
- The academic integrity policy will be strictly enforced. Each student must complete the exam **independently**, with no collaboration allowed.
- The total score for the exam is 100 points.

Name: ______________________________

UID: ______________________________

Part A: Multiple-Choice Problems (20 points)

1	2	3	4	5

6	7	8	9	10

1. **(2 points)** Which statement of three representative graph neural network architectures: Graph Attention Networks (GAT), Graph Convolutional Networks (GCN), and GraphSAGE is true?

    A. GAT uniquely employs fixed-weight neighborhood aggregation, optimizing computational efficiency for large graphs.

    B. Only GAT introduces an attention mechanism, allowing it to dynamically weight the importance of each neighbor's features for more accurate node representation.

    C. GraphSAGE is superior to GAT because it exclusively supports inductive learning, whereas GAT and GCN do not.

    D. GCN outperforms GAT in link prediction tasks by utilizing global pooling functions to aggregate neighborhood information more effectively.

    Answer: B

2. **(2 points)** Which statement best describes the common techniques in recommender systems?

    A. Matrix factorization that treats ALL missing values as 0 in the user-item interaction matrix is an effective approach for implicit feedback recommendation.

    B. Bayesian Personalized Ranking (BPR) aims at ranking the user's observed interactions over unobserved ones, using a pairwise ranking loss function.

    C. Implicit matrix factorization techniques are designed to reconstruct the original user-item matrix exactly, treating all non-interaction as lack of information rather than negative feedback.

    D. Both AUC and Precision@k rely on explicit ratings to compute, making them less effective for implicit feedback measures where such ratings are not available.

    Answer: B

3. **(2 points)** Given the applications of Graph Neural Networks (GNNs) for various tasks, identify which of the following statement regarding training objectives is not correct. Use $\boldsymbol{u}_i$ to denote embedding of node i, y_i for the label of node i, $\boldsymbol{g}$ for the graph embedding, y for the graph label, and $\boldsymbol{A}_{ij} = 1$ to indicate the presence of an edge between nodes i and j and 0 otherwise.

    A. For link prediction, the objective is to maximize the likelihood of correctly predicting $\boldsymbol{A}_{ij}$ based on the embeddings $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ of the two involved nodes.

    B. For node classification, the training objective is to minimize the cross entropy between the embedding of nodes and the label of the nodes.

    C. For graph classification, the training objective is to minimize the cross entropy between the predicted probability of graph belonging to each class, which is based on graph-level embedding, and their ground truth class label.

D. For graph similarity search, the training objective is to minimize the graph-embedding based similarity measure and the ground truth similarity measure between graphs.

Answer: B

4. **(2 points)** What key principle underpins the operation of Graph Convolutional Networks (GCNs), allowing them to effectively capture the structural features of the graph?

A. GCNs apply traditional Convolutional Neural Network (CNN) filters directly to the graph structure, treating graph nodes as regular grid points in Euclidean space.

B. GCNs employ a dynamic routing mechanism between nodes, where the convolution operation is guided by path lengths and node centrality measures to update node representations.

C. GCNs leverage an eigendecomposition of the graph Laplacian to apply convolution in the spectral domain, requiring the computation of the Laplacian for every convolution layer.

D. GCNs aggregate and transform neighbor node features through learnable weights, effectively capturing local graph topology by averaging features from a node's immediate neighbors.

Answer: D

5. **(2 points)** Which of the following statements accurately reflects the limitations or characteristics of knowledge graph embedding models in terms of their ability to model relationships, use of negative sampling, and their training objectives?

Approach	Scoring function
TransE	$f(h,r,t) = -\|\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}\|$
DistMult	$f(h,r,t) = \langle \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t} \rangle$
RotatE	$f(h,r,t) = -\|\boldsymbol{h} \circ \boldsymbol{r} - \boldsymbol{t}\|$, where $\circ$ denotes the element-wise product and $\boldsymbol{r}$ is represented in complex space to enable rotation

A. TransE struggles with modeling many-to-many relations, such as "GrandparentOf" relation, where a head entity can be associated with multiple tails and multiple heads can be associated with a tail entity .

B. DistMult, due to its symmetric scoring function, excels in modeling asymmetric relations, making it particularly useful for datasets where the directionality of relations significantly impacts the graph structure.

C. RotatE inherently models all types of relation patterns, including symmetry, antisymmetry, inversion, and composition, without any limitations, by utilizing complex numbers to represent relations.

D. Negative sampling is a technique uniquely beneficial to DistMult and not applicable to TransE and RotatE, as it helps in distinguishing between positive and negative examples only in symmetric relation models.

Answer: A

6. **(2 points)** Which of the following is NOT a proper regularization mechanism for deep neural networks?

A. Add additional weight penalty terms in the loss function, such as L-1 and L-2 norm of weights.

B. Dropout some of the weights when performing forward computation during training.

C. Early stopping for training based on validation.

D. Use gradient descent rather than stochatic gradient descent for optimization.

Answer: D

7. **(2 points)** Which of the following is INCORRECT about K-Means algorithm?

A. K-Means is a type of hard clustering algorithm, where each data point belongs to one exact cluster.

B. The K-Means algorithm can be initialized randomly.

C. Gradient decent can be used to optimize the objective function of K-Means.

D. It cannot handle clusters with non-spherical shape.

Answer: C

8. **(2 points)** Suppose you want to classify documents as either spam or non-spam and you decide to use a discriminative model for this task. Which of the following model is the one to choose?

A. Naive Bayes.

B. Logistic Regression.

C. pLSA.

D. Linear Regression.

Answer: B

9. **(2 points)** Which of the following is NOT a weakness of the Gaussian Mixture Model?

A. No guarantee of convergence to global optimal value.

B. Difficult to estimate the number of clusters.

C. Clusters require a large number of parameters to specify.

D. Can only deal with spherical clusters.

Answer: C

10. **(2 points)** Regarding Network Embedding algorithm known as LINE, which of the following is NOT TRUE?

A. It uses random walk on graphs to measure the similarity between nodes.

B. Its uses first and second order proximity to characterize node similarity.

C. The training minimizes the KL divergence between empirical link distribution and modeled link distribution.

D. It uses negative sampling for better learning efficiency.

Answer: A

Part B: Open-Answer Problems

Problem 11: Mixture Models for Binary Data (20 points)

Suppose we have a dataset of black and white images. The inputs $\boldsymbol{x}^{(i)}$ for the ith image are vectors of binary values corresponding to black and white pixel values, and the goal is to cluster the images into groups. For simplicity, you can use the answer from previous parts of the problem to later parts (e.g., use the answer in Part 1 for Part 2, etc.) For this problem, you simply need to consider a slight change to the Gaussian mixture model.

1. **(5 points)** Consider a binary random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_D) \in \{0,1\}^D$. What is a choice of probability distribution to model the observed pixel value for each dimension of this random vector? (e.g., model $P(x_d) = 1$, $P(x_d = 0)$ for $d \in \{1, \cdots, D\}$)

2. **(5 points)** Using the proposed probability distribution above, suppose that $P(x_d = 1) = p_d$, where x_d is the random variable for the dth dimension of $\boldsymbol{x}$. Assume that random variables at each dimension are independent with each other (i.e., x_d is independent of x_d' if $d \neq d'$); write down the expression for $P(x_d)$ and $P(\boldsymbol{x})$ using $p_1, \ldots, p_D$.

3. **(5 points)** Suppose that there are K clusters, where kth cluster is associated with distribution described in Part 2 with parameter $\boldsymbol{p}_k$ and p_{kd} denotes the d_{th} dimension in $\boldsymbol{p}_k$. The prior probability each data point belonging to each cluster is $p(z = k) = \pi_k$. Write down the expression for $P(\boldsymbol{x}^{(i)})$ following the mixture model assumption.

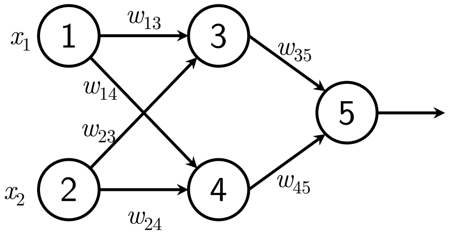
   (**Hint:** Compare it with the Gaussian mixture model and consider $P(\boldsymbol{x}^{(i)} \mid z = k)$ and $p(z = k)$.)

4. **(5 points)** Suppose that the dataset is generated i.i.d. $X = \{\boldsymbol{x}^{(i)}\}_{i=1,\ldots,n}$. Write down the expression for the log-likelihood of the dataset. You can use the solution in Part 3.

Answer:

Problem 12: Neural Network (26 points)

As a data scientist working for a hospital, you want to use neural network to assess the health of patients. Consider the fully-connected neural network given below, where input x_1, x_2 are real numbers, the output activation function is the sigmoid function and all the rest activation functions are ReLU.



The weights and biases are given in the following table.

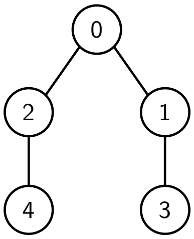
w_{13}	w_{14}	w_{23}	w_{24}	w_{35}	w_{45}	b_3	b_4	b_5
4	−3	4	−3	5	5	−2	5	−5

1. **(6 points)** Compute the model output when the input data is $(x_1, x_2) = (1, 1)$. You can leave the sigmoid function expression without evaluating the numerical value.
2. **(5 points)** What is the kind of supervised learning task this neural network is able to perform? State the reason.
3. **(5 points)** Suppose you want to use this neural network to predict the Cholesterol level (a categorical variable of $K > 2$ categories) of patients given a tabular datasets of real-valued features, what are the modifications you need to make to the network?
4. **(5 points)** Suppose you want to predict the Height (a real value) of patients. What are the modifications you need to make to the network? Which loss function do you like to use?
5. **(5 points)** A medical exam for the brain requires processing of MRI scans, which are images with potential regions of brain tumor. To determine if a user has a tumor, what are the possible modifications you can make to the current network? You are allowed to make major changes.

Answer:

Problem 13: Label Propagation (10 points)

Consider the following graph G, which consists of 5 nodes labeled from 0 to 4.



Suppose we initially have the following labels for each node:

- Node 0: Label A
- Node 1: Unlabeled
- Node 2: Label B
- Node 3: Unlabeled
- Node 4: Unlabeled

1. **(4 points)** Show the adjacency matrix of the graph G.
2. **(6 points)** Using a simple label propagation algorithm, where each unlabeled node adopts the label of the majority of its neighbors (or remains unlabeled if there is a tie), determine the labels of all nodes after one iteration of label propagation.

Answer:

Problem 14: Recommender Systems (12 points)

Consider the table of ratings below.

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	5		3		4
User 2	3	3	1	5	2
User 3		5	5	2	
User 4	3	5	1		1
User 5			5	2	5
User 6	5	3	?	5	

In this problem, 2 decimal places are sufficient for your answer, or you may give exact expressions.

1. **(4 points)** You wish to predict the rating of Item 3 by User 6. Suppose we use user-user collaborative filtering and use Pearson correlation as the similarity metric. What is the Pearson correlation coefficient between User 2 and User 6?

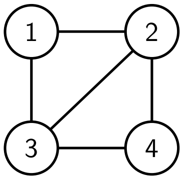
2. **(4 points)** Which other users could be used to help predict the rating of Item 3 by User 6?

3. **(4 points)** One problem with pure collaborative filtering is newcomers who have not rated anything. This is where the content-based approach comes in handy. Suppose we have a new user 7 with profile features $[1, 3, 0, 2, -2]$. We also know that the profile features for Item 3 are $[-1, 0, 2, 2, 0]$. What is the cosine similarity between User 7 and Item 3? You should give the value of the cosine itself, rather than the angle with that cosine.

Answer:

Problem 15: Spectral Clustering (12 points)

Apply spectral graph clustering to the following graph.



1. **(4 points)** Write the Laplacian matrix $\boldsymbol{L}$ for this graph. All the edges have weight 1.
2. **(4 points)** Consider the minimum bisection problem, where we find an indicator vector $\boldsymbol{y}$ that minimizes $\boldsymbol{y}^\top \boldsymbol{L}\boldsymbol{y}$, subject to the balance constraint $\mathbf{1}^\top \boldsymbol{y} = 0$ and the strict binary constraint $\forall i, y_i = 1$ or $y_i = -1$. Write an indicator vector $\boldsymbol{y}$ that represents a minimum bisection of this graph.
3. **(4 points)** Suppose we relax (discard) the binary constraint and replace it with the weaker constraint $\boldsymbol{y}^\top \boldsymbol{y} =$ constant, permitting $\boldsymbol{y}$ to have real-valued components. (We keep the balance constraint.) What indicator vector is a solution to the relaxed optimization problem? What is its eigenvalue?

   (**Hint:** Look at the symmetries of the graph. Given that the continuous values of the y_i's permit some of the vertices to be at or near zero, what symmetry do you think would minimize the continuous-valued cut? Guess and then check whether it is indeed an eigenvector.)

Answer: