Big Data Analytics
Stony Brook University
CSE538 - Spring 2023

Exam 2 Review Guide

Exam Time: Tuesday, 4/30, 4:00pm

Suggestions for study:

1. Review the readings in the syllabus schedule
2. Review your notes associated with the slides and topics below – use lecture recordings where unclear.
3. Review assignment 2 and quizzes
   (In Brightspace click Weekly Quiz -> title of the specific quiz -> in the quiz page you will see a down arrow besides the title of the quiz.)
4. Practice example questions below and exercises in the book.
   (only those related to the topics we've covered).
5. Ask questions on Piazza

Topics for the exam:

- Transformer Language Modeling
  - Key works in language modeling, vector semantics, and their combination (Markov, Shannon, Osgood, Deerwater, Collobert and Weston)
  - Artificial Neural Networks
    - activation functions
    - backpropogation
    - feedforward neural network (multilayer or deep neural network)
    - idea of recurrent neural network
  - Transformer Components
    - Self-Attention
      - attention scoring function
      - key, queries, values
    - Positional Embeddings
    - Multi-headed attention
    - Residual Connections, FFNs
    - Dimensionality of each component
  - LM Objectives
    - Masked Language Modeling (autoencoder)
    - Generative language modeling (autoregressive)
  - Encoder-Decoder (sequence to sequence models)
  - Generating Language: Greedy, Beam Search, Random Walk
  - Approaches to Apply

        - pretraining, continued pretraining (aka finetuning the LM)
        - task fine-tuning
        - instruction tuning
        - use as contextual embeddings
        - idea of RAG
        - few shot, zero-shot
- Ethics in NLP
    - Bias.
        - Disparities
            - outcome
            - error
        - Origins
            - selection bias
            - label bias
            - overamplification
            - semantic/embedding bias
    - Ethical Principles of the ACM
    - Research Ethical Considerations
- Human-Centered NLP
    - Differential Language Analysis
        - Linear Models (Correlation, multivariate standardized linear regression)
        - log-odds ratio with informative Dirichlet prior
    - User Factor Inclusion
        - Additive Inclusion (as "controls)
        - Adaptive Inclusion (user factor adaptation)
    - Human Language Modeling
        - Objective, in multiple forms
        - HaRT model
- Speech Processing
    - Concept: Automatic Speech Recognition (ASR)
    - Encoding Waves:
        - FFT outputs: Spectrograms
        - mel spectrograms
    - Wave2Vec Approach
    - Whisper Approach
    - Current Challenges
- Question Answering and Machine Translation
    - Reading Comprehension
    - Open Domain QA
    - Commonsense QA
    - Evaluations
    - Multi-hop QA
    - BLUE score

- TBD

Exam Day Procedure:

Students remain outside while exam is being setup. The door will open to the classroom approximately 5 minutes before start time.

Once entering, students should sit at a seat with an exam in front of it and with at least 1 empty seat between them and other students.

No items are permitted outside one's bag, backpack, or briefcase except:
(1) pen or pencil,
(2) SBU student id, and
(3) optionally, a bottle of water
No calculators or other materials are permitted on one's desk. The last page of the exam is scrap paper that can be torn off after the exam begins. This page must be turned in with a name on it.

Once the exam begins, put your name on every page of the exam.

Students will be given 5 minute and 1 minute warnings before the end of exam, which will be announced as "Exam is over. Pens and pencils down". **Once the exam ends, all exams must be closed and all writing utensils down. Students who do not close the exam or who do not put their pen down will receive a 50-point deduction off their exam grade.**

FAQ:

- **Will there be programming on the exam?**
  - Yes, as we have gone over and in the syllabus: Coding specific algorithms is a great way to both learn and demonstrate understand of concepts and processes.
  - A good rule of thumb is to just use python code, though pseudocode will be allowed.
  - As long as the key pieces of the algorithm are clear, small syntax errors won't count off (for example, missing a colon after a "for" will not count against you as long as the scope of the loop is clear. If you miss the colon and don't indent clearly, then it's unclear and can reduce points).
- **How much of the readings will be covered?**
  - The readings are considered required. To fully understand a topic, one should approach from multiple perspectives and the readings are different perspective, following the same concepts of the class. The readings also cover some concepts in more depth than we get to in class. Some of these are even pointed out in the slides. Therefore, the readings are very helpful toward mastery of the class concepts.
  - That said, typically, there will be around 1 question that covers material from the readings which wasn't covered as extensively in class.
- **Will the exam be cumulative?**
  - The exam will focus on the second half of the course (the topics above). However, note that many of them build on concepts from the first part, or one implements algorithms in the second part using dataframeworks from the first part.

Please post any other questions to Piazza by Sunday.

Example Questions.
Example Questions.

1. Circle "T" if the statement below is true, circle "F" if the statement is false, and leave blank if you are completly unsure.

A. **T / F** The hyperbolic tangent and rectified linear unit are sigmoid activation functions.

B. **T / F** Each of the *query*, *key*, and *value* of attention must reference different hidden state nodes or layers of a neural network.

C. **T / F** In sequence to sequence modeling, using an "encoder" means one uses a crytographic algorithm to assist with the translation.

D. **T / F** In visual QA one tries to answer a question in natural language where one or more image is used as context.

2. Pseudocode: . Complete the method below such that the cross-entropy loss is used when training the transformer language model, defined by "model", in pytorch. Assume "x" is input.

```
sgd = torch.optim.SGD(model.parameters(), lr=learning_rate)
#training loop:
    for i in range(epochs):
        model.train()
        sgd.zero_grad()
        #forward pass:

        y_pred =

        loss =

        #backward:
        loss.backward()
        sgd.step()
```

3. Short Answer. Both “selection bias” and “label bias” can originate by having biases in the outcomes of the training data. Please explain the difference between the two.

4. Multiple Choice. For speech processing, the spectrogram is is run through the mel filter, applying which function:

A. Polynomial
B. Square root
C. Logarithmic
D. GeLu
E. Exponential

5. problem solving: Assume you are applying attention with the following scoring function:

$$\psi_{dp}(h_i, s) = s^T h_i$$

Given:

$h_{i-1} = [1, -1]$, $h_i = [1, 1]$, $h_{i-1} = [-1, 0]$,
$s_{i-1} = [0, 1]$, $s_i = [-1, -1]$, $s_{i-1} = [1, 1]$,

Which s (i.e. value) would receive the most attention when given h_i as the query and s as the key? Justify your answer.

ANSWER KEY BELOW

1. Circle "T" if the statement below is true, circle "F" if the statement is false, and leave blank if you are completly unsure.

A. **T** / (**F**) The hyperbolic tangent and rectified linear unit are sigmoid activation functions.

B. **T** / (**F**) Each of the *query*, *key*, and *value* of attention must reference different hidden state nodes or layers of a neural network.

C. **T** / (**F**) In sequence to sequence modeling, using an "encoder" means one uses a crytographic algorithm to assist with the translation.

D. (**T**) / **F** In visual QA one tries to answer a question in natural language where one or more image is used as context.

2. (pseudocode). Complete the method below such that the cross-entropy loss is used when training the language model in pytorch. Assume "x" is input.

```
sgd = torch.optim.SGD(model.parameters(), lr=learning_rate)
#training loop:
    for i in range(epochs):
        model.train()
        sgd.zero_grad()
        #forward pass:
        y_pred = model(x)
        loss = torch.mean(-torch.sum(y*torch.log(y_pred))
```

```
#backward:
loss.backward()
sgd.step()
```

3. Short Answer. Both "selection bias" and "label bias" can originate by having biases in the outcomes of the training data. Please explain the difference between the two.

Selection bias occurs when the distribution of individuals for whome the outcomes are derived do not match the ideal distribution. Label bias occurs when annotators inject their own biases (perhaps inadvertently) in how they label the data (outcomes). Selection concerns the people the outcomes are about while label bias results in different labels than would be ideal.

4. Multiple Choice. For speech processing, the spectrogram is is run through the mel filter, applying which function:

A. Polynomial
B. Square root
C. Logarithmic
D. GeLu
E. Exponential

5. problem solving: Assume you are applying attention with the following scoring function:

$$\psi_{dp}(h_i, s) = s^T h_i$$

Given:

$h_{i-1} = [1, -1]$, $h_i = [1, 1]$, $h_{i-1} = [-1, 0]$,
$s_{i-1} = [0, 1]$, $s_i = [-1, -1]$, $s_{i-1} = [1, 1]$,

Which s (i.e. value) would receive the most attention when given h_i as the query and s as the key? Justify your answer.

S_{i+1} -- it has the largest dot product, which is how the scoring function is defined, with the query, h_i

$s_{i-1}^T h_i = 0 + 1 = 1$

$s_i^T h_i = -1 + -1 = -2$

$s_{i+1}^T h_i = 1 + 1 = 2$