## Quiz 3 A

**Question 1.** According to Lecture 16 on Machine Learning Fairness, which of these is NOT a type of bias that 'occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution'?

- (a) Selection Bias
- (b) Experimenter's Bias
- (c) Non-response Bias
- (d) Coverage Bias
- (e) Sampling Bias

**Question 2.** According to lecture 14, what is one disadvantage of using Bag-of-Words encoding technique to represent a sentence?

- (a) Bag-of-Words encoding technique requires a large amount of data to train its representation model.
- (b) Bag-of-Words encoding technique assigns lower weights to frequent words, making them less significant in a representation.
- (c) Bag-of-Words encoding technique does not consider word order in a sentence in its representation.
- (d) Bag-of-Words encoding technique applies lemmatization to every word in a sentence, even when it is not required.
- (e) Bag-of-Words encoding technique cannot be used to represent rare words.

**Question 3.** When working on time critical tasks in crowdsourcing, which of the following is a method to increase HIT performance, according to lecture15 crowdsourcing slides?

- (a) Recruiting a few expert Turkers for general HITs
- (b) Partiontioning human intelligence tasks evenly by worker ability.
- (c) Creating flash teams by paing users a retainer to be available.
- (d) Ensuring all workers have plenty of time to complete the HIT.
- (e) Maxmizing the allowable pay for each HIT

**Question 4.** Which of the following statements is least accurate when discussing fairness in ai pipelines?

- (a) Machine learning can be used but data engineers should avoid descriminative tasks.
- (b) Computed risk scores can be used as advisory measures for human decision making.
- (c) Removing sensitive attributes (e.g. race and gender) from the data does not mean race and gender are not used in the model.
- (d) Equal opportunity measures do not acknowledge harms from false positives.
- (e) The Impossible Theorem of Fairness states that no more than two out of three specific fairness criteria can be satisfied simultaneously by a well-calibrated classifier.

**Question 5.** According to the slides in Lecture 14 about text vectorization, what is the key difference between CBOW and Skip-gram models in Word2Vec?

- (a) CBOW uses context words to predict the middle word, while Skip-gram uses the middle word to predict surrounding ones.
- (b) CBOW performs matrix multiplication while Skip-gram uses softmax activation.
- (c) CBOW is better for rare words and needs more training data than Skip-gram.
- (d) CBOW requires higher dimensionality vectors than Skip-gram to achieve similar performance.
- (e) Skip-gram uses one-hot encoding while CBOW uses bag-of-words representation.

**Question 6.** According to Lecture 15, what is one method used to improve data quality in crowdsourced annotation tasks?

- (a) Let workers choose their own pay rates to encourage competition.
- (b) Use multiple annotations per task and rely on redundancy.
- (c) Randomly assign tasks without providing qualifications.
- (d) Rely on a single annotation to reduce overhead costs.
- (e) Penalize workers who complete tasks quickly.

**Question 7.** In the discussion of a market for lemons, which of the following statements is most accurate?

- (a) The buyer will take the minimum quality of all similar products in their decision as to how much to pay
- (b) Good workers leave the market because they are not compensated for their work.
- (c) The seller is incentivised to sell less overall products
- (d) The quality of the market increase for high end and low end goods.
- (e) The buyer will average the quality of all similar products in their decision as to how much to pay.

**Question 8.** According to Lecture 14 on Encodings, what is a key distinction between stemming and lemmatization in the process of text normalization?

- (a) Stemming requires a dictionary lookup, whereas lemmatization uses fixed regular expressions.
- (b) Lemmatization removes all stop words, while stemming expands contractions.
- (c) Stemming uses rule-based techniques, while lemmatization uses a dictionary to transform words to their grammatical base.
- (d) Lemmatization and stemming are identical processes that normalize words using statistical models.
- (e) Stemming is more accurate than lemmatization because it preserves the syntactic role of a word in a sentence.

**Question 9.** When using word analogies to find the relationship between words, which of the following statements is most accurate?

- (a) Analoges only hold for a small number of nouns.
- (b) In a large dataset misspellings change the meaning of embeddings.
- (c) The analogy is a non-linear relationship in the embedding space, but can be approximated by a linear relationship.
- (d) The analogy is a linear relationship in the embedding space.
- (e) For domain specific datasets the vector operations for computing analogies have less variance than general datasets.

**Question 10.** According to Lecture 12, which of the following is NOT a method used to measure the distance between clusters?

- (a) Ward's method
- (b) Complete Link
- (c) Skip-gram method
- (d) Centroids
- (e) Average Link

Question 11. According to lecture 12, what causes distortion in the k-means clustering method?

- (a) Its desire to sort data based on cuteness.
- (b) Its preference for non-numeric data.
- (c) Sensitivity to outliers.
- (d) The use of medoids instead of means.
- (e) The need for labeled training data.

Question 12. According to Lecture 16 on Machine Learning Fairness, what best describes automation bias in the context of AI/ML systems?

- (a) A tendency to prefer decisions from automated systems even when they are flawed or have higher error rates.
- (b) A tendency to exclude automation tools from the model evaluation process due to their perceived unreliability.
- (c) A tendency of automation tools to reinforce sampling bias during data collection.
- $\left( d\right)$  A situation in which robots start making decisions without asking for human permission.
- (e) A bias that arises when automated systems overfit to in-group data at the expense of out-group generalization.

**Question 13.** According to the encodings lecture, which of the following opinions are most accurate concerning learned terms in the vector embedding space.

- (a) The magnitude of the terms changes its meaning.
- (b) Terms that are used and occur in the same context tend to purport the same meaning.
- (c) Location of embeddings is more important than the direction.
- (d) Stop words do not appear in the embedding space.
- (e) Using Cosine similarity means that location does not depend on the frequency of the term.

Question 14. According to lecture 16, what is Simpson's Paradox?

- (a) When the r coefficient is between 0.3 and 0.7.
- (b) When a trend between two variables is reversed in all subgroups of the data.
- (c) When a trend between two variables is reversed for some subgroups.
- (d) When model builders unconsciously process data in ways that affirm preexisting beliefs and hypotheses.
- (e) When the r coefficient is between 0 and 0.3.

Question 15. According to Lecture 11, what is the primary advantage of using PyTorch Lightning over standard PyTorch when developing neural network models?

- (a) It automatically selects the optimal neural network architecture for any given dataset.
- (b) It simplifies the training loop code by abstracting away engineering complexity while maintaining full PyTorch compatibility.
- (c) It eliminates the need for GPUs by optimizing CPU-based computations.
- $\rm (d)~$  It only works with pre-defined models and doesn't support custom neural network architectures.
- (e) It requires more code to implement but provides better performance than standard PyTorch.

Question 16. According to lecture 14 on encodings, if we have the following vocabulary: dog (1), cat (2), person (3), holding (4), tree (5), computer (6), and using (7), what would be the bag-of-words representation for the phrase 'computer person holding cat'?

(a) {6, 3, 4, 2} [0, 1, 1, 1, 0, 1, 0]
(b) {3, 4, 2} [0, 1, 1, 1, 0, 0, 0]
(c) {6, 3, 4, 1} [1, 0, 1, 1, 0, 1, 0]
(d) {6, 3, 4, 6} [0, 0, 1, 1, 0, 2, 0]
(e) {6, 3, 7, 6} [0, 0, 1, 0, 0, 2, 1]

**Question 17.** According to lecture 14, what is one advantage of using SpaCy embeddings over TF-IDF for text vectorization?

- (a) SpaCy embeddings randomly shuffle word vectors to improve generalization.
- $\rm (b)~$  TF-IDF embeddings are context-aware, while SpaCy embeddings treat each word in isolation.
- (c) SpaCy embeddings are derived using inverse document frequency and do not require any pretrained models.
- (d) SpaCy embeddings capture semantic meaning and word context, unlike TF-IDF which is based purely on word frequency.
- (e) SpaCy embeddings are stored as sparse matrices which reduce memory usage compared to TF-IDF.

**Question 18.** As per the lecture14 slides on encodings, what is the primary function of Softmax layer in standard Word2Vec models?

- (a) It converts the model output to the probability distribution of vocabulary words
- (b) It acts as the loss function while model training
- (c) It generates the final dense word embeddings for the vocabulary words
- (d) It normalizes the embedding vectors
- (e) It defines the context window size

**Question 19.** When examining the pipeline for the presence of bias with of the following approaches is considered a best practice?

- (a) Increasing the scale of the data to include more synthetic data to help move the fairness metrics thresholds.
- (b) Using select fairness metrics to remove bias from deployed systems.
- (c) Selecting the key fairness metric and recording its value in logs.
- (d) Ensuring fairness metrics are met and prioritized over accuracy metrics.
- (e) Stopping the pipeline completely before a bias audit is performed.

Question 20. According to lecture 15, which of the following is a significant challenge in ensuring fair compensation for workers on crowdsourcing platforms like Amazon Mechanical Turk?

- (a) The restriction of crowdsourcing platforms to specific geographic regions.
- (b) The requirement for workers to complete extensive training before joining.
- (c) The lack of standardized pricing for tasks, leading to inconsistent pay rates.
- (d) The inability to assign qualifications to workers, reducing task quality.
- (e) The absence of a system for monitoring task completion times.

UNIVERSITY OF FLORIDA, COMPUTER INFORMATION SCIENCE & ENGINEERING