cis6930sp25
March 7, 2025

Quiz 2 C

Question 1. According to lecture 8, which statement best describes a key property of Merge-Center Clustering that distinguishes it from Center Clustering?

(a) `When two clusters merge, the resulting cluster can have multiple center nodes instead of just one.`

(b) `Merged clusters always discard their original centers and elect a completely new center node.`

(c) `It only merges clusters if they have the exact same single center node to begin with.`

(d) `Merge-Center Clustering must repeatedly scan the entire list of pairs to recalculate center nodes.`

(e) `This approach never requires sorting the list of similar pairs prior to merging clusters.`

Question 2. According to lecture 10, What does a Receiver Operating Characteristic (ROC) Curve primarily show?

(a) `The tradeoff between True Positive Rate and False Positive Rate.`

(b) `The relationship between precision and recall.`

(c) `The absolute difference between predicted rate and actual values in a regression task.`

(d) `The fraction of correct predictions over total predictions.`

(e) `The proportion of correctly classified positive instances among predicted positives .`

Question 3. According to Lecture 9, which of the following is not true regarding Entity Resolution?

(a) `Entity resolution in text is the process of identifying and clustering different manifestations of the same real world object.`

(b) `Factors across entities are called repeat factors.`

(c) `Factors within entities are called attract factors.`

(d) `It is a necessary pre-step for advanced stages of the data pipeline.`

(e) Main Difficulty in Entity Resolution is Because of ambiguity

Question 4. According to lecture 8, What is the primary goal of block processing in entity resolution?

(a) To compare records within blocks to find matches based on similarity.

(b) To sort records by specific field values and use a sliding window for comparison.

(c) To refine blocks and minimize the number of comparisons.

(d) To tokenize attribute values and create blocks based on tokens.

(e) To group matched records into entities.

Question 5. According to Lecture 10, What happens when the classification threshold is made higher in a binary classification task?

(a) Both Precision and Recall increase.

(b) Precision decreases, and Recall increases.

(c) Precision remains constant, while Recall decreases.

(d) Both Precision and Recall decrease.

(e) Precision increases, and Recall decreases.

Question 6. According to lecture 11, Which of the following is NOT true about tensors in pytorch?

(a) Tensors attributes describe their shape, datatype, and the device.

(b) Pytorch and other libraries are built around manipulating, and processing large data in tensors efficiently.

(c) Tensors is a data structure analogous to a numpy array(1-D) or matrix(2-D), but can represent even higher dimensions.

(d) Tensors infer the shape and datatype of the right hand side, making it convenient for loading data.

(e) The syntax for the dot product of two multidimensional tensors is A * B' .

Question 7. If SS_res = Sum of squared residuals and SS_tot = Total sum of squares,

According to the lecture on Machine learning metrics, in which case the R-squared defined as $\mathrm{R}^2 = 1 - \frac{SS\_res}{SS\_tot}$, will always be negative?

(a) `Model consistently predicting values farther than the mean of the values`

(b) `Model predicts values same as the mean`

(c) `Model predicts positive linear but weak relationship with the actual values`

(d) `R-squared can never be negative`

(e) `Model always underestimate the prediction but the residue decrease as the value increase`

Question 8. According to the lecture 11, what is the primary objective of Linear Regression in supervised learning?

(a) `To transform input variables using Scaling.`

(b) `To increase the precision of the model's predictions.`

(c) `To separate classes using a linear function.`

(d) `To maximize the accuracy of classification.`

(e) `To minimize the differences between the predicted and actual values.`

Question 9. According to Lecture 11, why do we divide data into training, validation, and test sets, and what is the actual need for validation data instead of directly using test data?

(a) `The validation set is unnecessary, and the model should be directly tested on the test set after training.`

(b) `Validation data is only used when there is not enough test data available.`

(c) `Validation data is used to fine-tune model parameters and prevent overfitting before evaluating on the test data.`

(d) `The training set alone is sufficient for model evaluation, making validation and test sets redundant.`

(e) `The test set is used for hyperparameter tuning, while the validation set is used for final evaluation.`

Question 10. According to lecture 9, why does the baseline entity resolution approach using Metropolis-Hastings sometimes reject a proposed merge?

(a) `Because once a mention is assigned to an entity, it cannot be reassigned.`

(b) `Because the proposed merge does not improve the overall entity resolution state.`

(c) `Because entity resolution does not allow probabilistic sampling.`

(d) `Because the algorithm requires all mentions to be merged in a single step.`

(e) `Because merges are only accepted if they involve identical strings.`

Question 11. According to lecture 10, in a Precision-Recall (PR) curve, how does the shape of the curve typically change when the dataset becomes more imbalanced (with far fewer positive examples compared to negative ones)?

(a) `The curve becomes steeper at the beginning but flatter at higher recall values.`

(b) `The curve shifts downward, indicating lower precision at all recall levels.`

(c) `The PR curve gets replaced by an ROC curve in case of imbalance.`

(d) `The shape of the curve does not change because the PR curve is independent of class distribution.`

(e) `The curve shifts upward, indicating higher precision at all recall levels.`

Question 12. According to Lecture 9, what is one key advantage of using probabilistic matching over deterministic matching in entity resolution?

(a) `Probabilistic matching uses quantum mechanics to resolve entities faster than traditional methods.`

(b) `Probabilistic matching guarantees perfect matches for all records in the dataset.`

(c) `Probabilistic matching only works when all data fields are complete and standardized.`

(d) `Probabilistic matching accounts for uncertainty by assigning weights to attribute comparisons, improving accuracy in noisy datasets.`

(e) `Probabilistic matching eliminates the need for any predefined rules or logic.`

Question 13. In lecture 9 we have learned that entity resolution is the process of identifying and clustering different manifestations of the same entity. What is implied about the relationship between different manifestations?

(a) `They are always in different languages.`

(b) `They are always in different data formats.`

(c) `They refer to the same real-world object.`

(d) `They are always identical.`

(e) `They have no relation to each other.`

Question 14. According to Lecture 9: Entity Resolution, what is one advantage of using the Metropolis-Hastings algorithm in entity resolution?

(a) `It eliminates all ambiguity in entity resolution.`

(b) `It should be able to find a global optimum.`

(c) `It assigns all mentions to a single entity immediately.`

(d) `It ensures an exact solution to the entity resolution problem.`

(e) `It requires no proposal function for new entity assignments.`

Question 15. According to the lecture 10 slides and what was discussed in class, what is the key purpose of the transform() method in Scikit-learn's pipeline?

(a) `It is used to train a machine learning model on training data.`

(b) `It combines different transformers into a single preprocessing step.`

(c) `It scales feature values using min-max normalization.`

(d) `It takes feature inputs and returns predictions for the target variable.`

(e) `It applies operations to all variables in an input matrix.`

Question 16. According to lecture 10, what does Precision measure?

(a) `Out of all the possible positives, how many the model correctly identified.`

(b) `When the model predicted positive, how often it was right.`

(c) `The overall accuracy of the model across all classes.`

(d) `The square root of the recall score.`

(e) `The ability of the model to minimize false negatives.`

Question 17. According to lecture 11 and the discussion in class, given this line of code, what is the purpose of setting shuffle to True?

```
train_dataloader = DataLoader(training_data, batch_size=64, shuffle=True)
```

(a) `To prevent the model from learning the order of the data.`

(b) `To order the data in a specific way (e.g. alphabetically or chronologically) such that the model can be trained more efficiently and faster.`

(c) `To vary the batch sizes`

(d) `To keep the original order of the data`

(e) `To shuffle the tokens in text embeddings`

Question 18. According to lecture 10, what does the R-squared score represent in a linear regression model ?

(a) `It quantifies the percentage of correct predictions made by the model.`

(b) `It represents the proportion of the variance in the dependent variable that is explained by the model.`

(c) `It measures the average squared difference between the predicted and actual values.`

(d) `It calculates the mean absolute error between predictions and actual observations.`

(e) `It represents the degree of multicollinearity among the independent variables.`

Question 19. According to lecture 8, which of the following is NOT true about blocking

(a) Sorted neighborhood is a blocking method that sorts records into alphabetical order and utilizes and fixed-size window to generate a list of candidate pairs

(b) Meta-blocking transforms a block collection into a graph, where each node corresponds to a record, and edges link every pair of records that co-occur with a block.

(c) Cullotta blocking is a blocking method that utilizes a pre-trained token matrix to assign similarity scores to data

(d) Token blocking is a blocking method that partitions data into tokens typically of single worlds or small n-grams

(e) Standard blocking is an easy to implement blocking method that partitions data into blocks based on a specific attribute

Question 20. According to lecture 8 dealing with Entity Resolution, what is the primary aim of blocking in entity resolution?

(a) Compare records and identify matches.

(b) Partition the dataset into blocks based on a specific attribute.

(c) Extract n-grams from the text.

(d) Reduces the search space to identify the same entity.

(e) Refines blocks to eliminate unnecessary comparisons.

Question 21. According to lecture 9 on Entity Resolution, what is a key challenge when using Metropolis-Hastings for entity resolution in large datasets?

(a) The algorithm can take a very large number of samples to converge to an optimal solution.

(b) The method is deterministic and does not account for uncertainty in entity resolution.

(c) The Metropolis-Hastings algorithm does not support probabilistic sampling.

(d) Once an entity is assigned, it cannot be reassigned in subsequent iterations.

(e) It requires exact matches between entity mentions to function correctly.

Question 22. According to Lecture 7, what is one key advantage of the “Global as View” (GAV) approach compared to the “Local as View” (LAV) approach in data integration?

(a) `GAV eliminates the need for schema alignment between data sources.`

(b) `GAV ensures that all queries are executed directly on the original data sources.`

(c) `GAV is simpler to implement and allows control over the mediator’s behavior.`

(d) `GAV is a form of Artificial Intelligence.`

(e) `GAV automatically adapts to new data sources without additional configuration.`

Question 23. In lecture 11, we discussed the PyTorch utility library “DataLoader”. What do we receive while iterating through a DataLoader?

(a) `The entire dataset as a single tensor.`

(b) `Metadata about the dataset.`

(c) `Individual samples one at a time.`

(d) `Batches of samples.`

(e) `A data augmentation pipeline that transforms samples on the fly.`

Question 24. According to lecture 10, which of the following describes the purpose of a linear regression model?

(a) `To measure the tradeoff between precision and recall in classification tasks`

(b) `To evaluate a model’s ability to classify positive and negative cases using an ROC curve`

(c) `To fit a continuous line that minimizes the difference between predicted and actual values.`

(d) `To separate data points into distinct classes using a linear function`

(e) `To preprocess and transform data using Scikit-Learn’s API before model training`

Question 25. According to the lecture 10, what does Recall measure in a classification model?

(a) The sum of precision and specificity.

(b) The fraction of predicted positive cases that were actually correct.

(c) The fraction of actual positive cases that were correctly predicted by the model.

(d) The likelihood that a model will always predict positive.

(e) The total number of false positives in the model.

Question 26. According to Lecture 10, which of the following best describes the role of regression metrics in evaluating models?

(a) They optimize the hyperparameters of a model to minimize overfitting.

(b) They determine how well a classification model separates different classes.

(c) They analyze the distribution of categorical data within a dataset.

(d) They evaluate the efficiency of a machine learning model by assessing training time
.

(e) They measure the performance of a regression model by comparing predicted and actual values.

Question 27. According to lecture 10, which metric quantifies the proportion of correctly predicted positive instances among all instances predicted as positive?

(a) Accuracy.

(b) Recall.

(c) F-score.

(d) Precision.

(e) Area Under the Curve (AUC).

Question 28. Which statement about the relationship between PR curves and ROC curves is MOST accurate?

(a) `ROC curves cannot be used for multi-class classification problems, but PR curves can`

(b) `PR curves focus on the positive class, while ROC curves consider both classes`

(c) `PR curves are only useful for regression problems, while ROC curves are only for classification`

(d) `The area under the PR curve is always larger than the area under the ROC curve`

(e) `A model that performs well on ROC will always perform poorly on PR`

Question 29. According to lecture 10, which of the following scenarios would maximizing Recall be more important than maximizing Precision?

(a) `Building a fraud detection system where it's critical to identify as many fraudulent transactions as possible.`

(b) `Developing a system to identify high-value customers for a personalized marketing campaign, where resources are limited and it's crucial to only target the most likely candidates.`

(c) `Creating a news aggregator that aims to provide users with only the most relevant and accurate articles on a specific topic.`

(d) `Implementing a system to predict equipment failure in a manufacturing plant, where the primary goal is to minimize false alarms that would lead to unnecessary maintenance checks.`

(e) `Determining which model to use for predicting stock market trends.`

Question 30. According to lecture 10, what is a key reason why Mean Squared Error (MSE) is often preferred over Mean Absolute Error (MAE) in regression tasks?

(a) `MSE is always a better metric than MAE for evaluating regression models, regardless of context.`

(b) `MSE penalizes larger errors more heavily, making it more sensitive to outliers compared to MAE.`

(c) `MSE and MAE always produce the same ranking of models, so either can be used interchangeably.`

(d) `MSE is computationally less expensive than MAE because it avoids absolute value calculations.`

(e) `MSE measures classification accuracy, making it suitable for both regression and classification tasks.`

Question 31. According to Lecture 9: Entity Resolution, what is one key advantage of using factor graphs in entity resolution models?

(a) `Factor graphs are not real just like birds.`

(b) `They do not require statistical methods for inference.`

(c) `They eliminate ambiguity entirely.`

(d) `They can capture arbitrary relationships between random variables.`

(e) `They only work for small datasets.`

Question 32. According to Lecture 8, why is standardizing text (e.g., converting to lowercase and removing punctuation) a crucial preprocessing step in entity resolution?

(a) `It minimizes superficial differences so that similar records are more likely to be matched accurately.`

(b) `It completely eliminates the need for fuzzy matching.`

(c) `It reduces the computational cost by significantly downsizing the dataset.`

(d) `It automatically assigns higher weights to certain attributes.`

(e) `It increases the number of unique tokens, enhancing detail.`

Question 33. According to lecture 10, what is the primary advantage of using a Precision-Recall (PR) curve over a Receiver Operating Characteristic (ROC) curve when evaluating a classifier on an imbalanced dataset?

(a) `PR curves are more sensitive to class imbalance and provide a better representation of model performance on the minority class.`

(b) `PR curves can handle multi-class classification problems, while ROC curves cannot.`

(c) `PR curves always have a larger area under the curve compared to ROC curves.`

(d) `PR curves directly show the trade-off between precision and recall, while ROC curves do not consider precision.`

(e) `PR curves are computationally less expensive to calculate than ROC curves.`