cis6930sp25
March 7, 2025

Quiz 2 E

Question 1. 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 have no relation to each other.`

(b) `They are always identical.`

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

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

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

Question 2. According to lecture 8 dealing with Entity Resolution, what is the primary aim of blocking in entity resolution?

(a) `Reduces the search space to identify the same entity.`

(b) `Extract n-grams from the text.`

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

(d) `Refines blocks to eliminate unnecessary comparisons.`

(e) `Compare records and identify matches.`

Question 3. 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 4. According to lecture 10, which statement is correct about Receiver Operating Characteristic (ROC) Curve?

(a) `ROC curve is more sensitive to class imbalance than Precision-Recall curve`

(b) `ROC curve refers to the area under the curve`

(c) `ROC curve only uses False Positive Rate`

(d) `ROC curve's shape changes when a model changes the way it classifies only one outcome`

(e) `ROC curve shows how well a model can classify binary outputs`

Question 5. According to Lecture 10, which of the following best describes the role of regression metrics in evaluating models?

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

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

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

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

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

Question 6. 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 are only useful for regression problems, while ROC curves are only for classification`

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

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

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

Question 7. According to the lecture 10, why is accuracy not always a good metric for evaluating classification models, especially when dealing with imbalanced datasets?

(a) `Because accuracy can be high even when the model fails to correctly classify the minority class.`

(b) `Because accuracy measures how well a model memorizes the training data.`

(c) `Because accuracy is a probabilistic metric rather than a deterministic one.`

(d) `Because accuracy takes both precision and recall into account, making it unreliable .`

(e) `Because accuracy is only useful for linear models and not for non-linear models.`

Question 8. 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 test set is used for hyperparameter tuning, while the validation set is used for final evaluation.`

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

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

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

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

Question 9. According to lecture 10, what is the correct equation to solve for Precision?

(a) `FP / (FP + TP)`

(b) `TP / (TP + FN)`

(c) `(TP + TN) / (TP + TN + FP + FN)`

(d) `TP / (TP + FP)`

(e) `(2 * TP) / (2 * TP + FP + FN)`

Question 10. 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 $R^2 = 1 - \frac{SS\_res}{SS\_tot}$, will always be negative?

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

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

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

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

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

Question 11. According to lecture 10, what does Precision measure?

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

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

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

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

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

Question 12. 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 always have a larger area under the curve compared to ROC curves.`

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

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

(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.`

Question 13. According to the lecture 10 slides and what was discussed in class, what is the key distinction between regression and classification tasks in machine learning?

(a) `Regression is evaluated using precision and recall while classification uses MSE and RMSE`

(b) `Regression uses the fit() method while classification uses predict()`

(c) `Regression uses linear models while classification uses non-linear models`

(d) `Regression predicts continuous values while classification assigns discrete classes`

(e) `Regression requires feature scaling while classification doesn't`

Question 14. 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 shifts downward, indicating lower precision at all recall levels.`

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

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

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

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

Question 15. According to lecture 10, What does a Receiver Operating Characteristic (ROC) Curve primarily show?

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

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

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

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

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

Question 16. According to lecture 10, what does the R-squared score represent in a linear regression model ?

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

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

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

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

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

Question 17. According to Lecture 9: Entity Resolution, what is one advantage of using the Metropolis-Hastings algorithm in entity resolution?

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

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

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

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

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

Question 18. In lecture 11, we discussed the PyTorch utility library "DataLoader". What do we receive while iterating through a DataLoader?

(a) `Batches of samples.`

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

(c) `Metadata about the dataset.`

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

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

Question 19. According to lecture 8, which statement best explains why the Center Clustering algorithm requires sorting the list of similar pairs in descending order of similarity scores before performing a single scan?

(a) `It's sorted that way only to create alphabetical clusters based on node labels.`

(b) `It reverses the clustering order so that less similar pairs form the initial cluster centers.`

(c) `It ensures that the most similar nodes are clustered first, allowing the algorithm to identify centers among highly similar pairs early in the process.`

(d) `It guarantees that every node is scanned multiple times to refine its cluster assignment.`

(e) `It aligns with a requirement from a different clustering algorithm that mandates ascending order.`

Question 20. 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 21. According to lecture 8, What is the primary goal of block processing in entity resolution?

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

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

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

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

(e) `To group matched records into entities.`

Question 22. According to lecture 7 on Information Integration, if a self-driving car relies on multiple data sources (e.g., GPS, traffic cameras, weather reports), what is the biggest risk if integration is not handled properly?

(a) `The car developing emotions and refusing to follow traffic laws.`

(b) `Conflicting or inconsistent data leading to incorrect decisions.`

(c) `All data sources merging seamlessly without any errors, too quickly.`

(d) `The car refusing to start due to too much data.`

(e) `The car becoming fully autonomous without needing integration.`

Question 23. 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 is simpler to implement and allows control over the mediator's behavior.`

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

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

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

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

Question 24. According to lecture 11, Which of the following is NOT true about tensors in pytorch?

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

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

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

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

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

Question 25. According to lecture 9 on Entity Resolution, what is the primary challenge in resolving entity mentions across different text sources?

(a) `The inability of computers to process named entities in large datasets.`

(b) `The need to manually verify every entity mention for correctness.`

(c) `The redundancy of using both coreference resolution and entity resolution in the same system.`

(d) `The lack of factor graphs in modern entity resolution models.`

(e) `Ambiguity in entity mentions, where the same name may refer to different individuals or different names may refer to the same individual.`

Question 26. According to Lecture 10, what is one of the key reasons spaCy may be preferred over other NLP libraries like NLTK for production environments?

(a) `SpaCy is implemented in Java, making it compatible with non-Python systems.`

(b) `SpaCy uses pre-trained transformer models exclusively, unlike NLTK.`

(c) `SpaCy is optimized for speed and composability, leveraging Python integration to efficiently handle large-scale NLP tasks.`

(d) `SpaCy relies exclusively on rule-based methods, which are faster than statistical models.`

(e) `SpaCy is designed only for academic research, not production-level applications.`

Question 27. According to lecture 9 on Entity Resolution, what is a key challenge when using Metropolis-Hastings for entity resolution in large datasets?

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

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

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

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

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

Question 28. According to Lecture 9, what is the most effective way entity resolution differentiates between Jimmy Fallon and Jimmy Kimmel?

(a) `The model treats all mentions of "Jimmy" as the same entity until manually corrected.`

(b) `Uses repel factors to distinguish between different entities, ensuring that mentions of Jimmy Fallon and Jimmy Kimmel do not get clustered together.`

(c) `If two names share a common first name, they must belong to the same person.`

(d) `Rely solely on the spelling of names to tell apart different individuals.`

(e) `The entity resolution process randomly assigns mentions of people to different entities without considering context.`

Question 29. According to lecture 8, which statement best describes a key property of Merge-Center Clustering that distinguishes it from Center Clustering?

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

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

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

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

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

Question 30. According to lecture 10, which of the following statements accurately describes the components of a confusion matrix in a binary classification task?

(a) `False Negatives (FN) occur when positive instances are incorrectly classified as negative.`

(b) `True Positives (TP) represent the number of negative instances that are correctly classified as negative.`

(c) `True Positives (TP) are the instances where the model incorrectly predicts positive instances as negative.`

(d) `False Positives (FP) represent the number of positive instances incorrectly classified as negative.`

(e) `True Negatives (TN) refer to positive instances that are correctly classified as positive.`

Question 31. According to the lecture, which metric is most appropriate for evaluating a classification model when dealing with an imbalanced dataset?

(a) `Precision-Recall (PR) Curve.`

(b) `Root Mean Squared Error (RMSE).`

(c) `The number of correctly predicted samples.`

(d) `Accuracy.`

(e) `Mean Squared Error (MSE).`

Question 32. According to lecture 8, which of the following best describes the primary benefit associated with token blocking?

(a) `It requires significantly fewer comparisons than other blocking methods.`

(b) `It achieves high precision at the expense of recall.`

(c) `It achieves high recall but may reduce precision due to redundant block membership.`

(d) `It eliminates the need for any subsequent matching step.`

(e) `It groups records solely by exact attribute matches.`

Question 33. According to lecture 10, what does the term False Positive indicate in a binary classification task?

(a) `It indicates the total number of instances predicted as positive, regardless of correctness.`

(b) `It indicates the number of instances where the model correctly predicts the positive class.`

(c) `It indicates the number of instances where the model incorrectly predicts the negative class when the actual class is positive.`

(d) `It indicates the number of instances where the model correctly predicts the negative class.`

(e) `It indicates the number of instances where the model incorrectly predicts the positive class when the actual class is negative.`