cis6930sp25
March 7, 2025

Quiz 2 A

Question 1. 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 is a probabilistic metric rather than a deterministic one.`

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

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

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

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

Question 2. 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 automatically assigns higher weights to certain attributes.`

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

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

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

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

Question 3. According to the lecture, which metric is most appropriate for evaluating a classification model when dealing with an imbalanced dataset?

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

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

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

(d) `Accuracy.`

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

Question 4. According to the lecture 11, what is the primary objective of Linear Regression in supervised learning?

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

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

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

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

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

Question 5. 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 ensures that all queries are executed directly on the original data sources.`

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

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

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

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

Question 6. According to lecture 10, which statement is correct about Receiver Operating Characteristic (ROC) Curve?

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

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

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

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

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

Question 7. 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 group matched records into entities.`

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

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

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

Question 8. 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 identical.`

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

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

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

Question 9. 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 correctly predicts the negative class.`

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

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

Question 10. According to lecture 9 on Entity Resolution, what is a key challenge when using Metropolis-Hastings for entity resolution in large datasets?

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

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

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

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

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

Question 11. According to Lecture 10, What happens when the classification threshold is made higher in a binary classification task?

(a) `Precision increases, and Recall decreases.`

(b) `Precision remains constant, while Recall decreases.`

(c) `Precision decreases, and Recall increases.`

(d) `Both Precision and Recall increase.`

(e) `Both Precision and Recall decrease.`

Question 12. 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 positive linear but weak relationship with the actual values`

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

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

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

Question 13. According to Lecture 10, which of the following best describes the role of regression metrics in evaluating models?

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

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

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

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

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

Question 14. According to lecture 8 on cluster evaluation, which metric is considered the best for evaluating cluster quality while correcting for chance?

(a) Rand Index.

(b) Silhouette Score.

(c) Random Correlated Index.

(d) Adjusted Rand Index.

(e) Elbow Method.

Question 15. According to lecture 11, what is the function of a Pytorch DataLoader?

(a) Load a dataset from an external source into local storage.

(b) Iterate through a dataset, with each iteration consisting of a batch of train features and labels.

(c) Use a generative model to create a simulated dataset.

(d) Augment an existing dataset with techniques like perturbation.

Question 16. According to lecture 10 (slide 26), what is the formula to calculate accuracy?

(a) `TP / TP + FN.`

(b) `TP / TP + FP.`

(c) `1 - SSres / SSTot.`

(d) `2 * TP /  2 * TP + FP + FN.`

(e) `TP + TN / TP + TN + FP + FN.`

Question 17. According to the lecture 10, what does Recall measure in a classification model?

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

(b) `The sum of precision and specificity.`

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

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

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

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) `Metadata about the dataset.`

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

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

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

Question 19. 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 requires feature scaling while classification doesn't`

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

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

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

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

Question 20. According to lecture 10, what is the purpose of the `make_regression` function in Scikit-learn API?

(a) `This function loads a built-in dataset suitable for a regression problem.`

(b) `This function generates a random regression problem according to the arguments passed on to it.`

(c) `This function takes features as input and returns a value for the target variable from a regression model.`

(d) `This function preprocesses values using scaling before training a regression model.`

(e) `This function trains a regression model on the features and targets passed on to it .`

Question 21. According to lecture 9 on Entity Resolution, what is the primary challenge in resolving entity mentions across different text sources?

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

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

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

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

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

Question 22. According to lecture 10, what is the correct equation to solve for Precision?

(a) `TP / (TP + FN)`

(b) `FP / (FP + TP)`

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

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

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

Question 23. 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 PR curve gets replaced by an ROC curve in case of imbalance.`

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

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

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

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

Question 24. 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 penalizes larger errors more heavily, making it more sensitive to outliers compared to MAE.`

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

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

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

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

Question 25. 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 can handle multi-class classification problems, while ROC curves cannot.

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

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

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

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

Question 26. According to lecture 11, Which of the following is NOT true about tensors in pytorch?

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

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

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

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

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

Question 27. 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 designed only for academic research, not production-level applications.

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

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

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

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

Question 28. According to lecture 9, why does the baseline entity resolution approach using Metropolis-Hastings sometimes reject a proposed merge?

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

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

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

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

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

Question 29. 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 scales feature values using min-max normalization.`

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

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

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

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

Question 30. According to Lecture 9, which of the following is not true regarding Entity Resolution?

(a) `Main Difficulty in Entity Resolution is Because of ambiguity`

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

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

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

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

Question 31. According to lecture 8 dealing with Entity Resolution, what is the primary aim of blocking in entity resolution?

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

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

(c) `Compare records and identify matches.`

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

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

Question 32. 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 aligns with a requirement from a different clustering algorithm that mandates ascending order.`

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

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

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

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

Question 33. According to Lecture 9: Entity Resolution, what is one advantage of using the Metropolis-Hastings algorithm in entity resolution?

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

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

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

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

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