cis6930sp25
March 7, 2025

Quiz 2 B

Question 1. According to lecture 8 on cluster evaluation, which metric is considered the best for evaluating cluster quality while correcting for chance?

(a) `Random Correlated Index.`

(b) `Rand Index.`

(c) `Adjusted Rand Index.`

(d) `Elbow Method.`

(e) `Silhouette Score.`

Question 2. According to Lecture 10, What happens when the classification threshold is made higher in a binary classification task?

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

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

(c) `Both Precision and Recall increase.`

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

(e) `Both Precision and Recall decrease.`

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 automatically assigns higher weights to certain attributes.`

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

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

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

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

(b) `Batches of samples.`

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

(d) `Metadata about the dataset.`

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

Question 5. 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 refusing to start due to too much data.`

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

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

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

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

Question 6. According to Lecture 9, what is one key advantage of using probabilistic matching over deterministic matching in entity resolution?

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

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

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

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

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

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 always underestimate the prediction but the residue decrease as the value increase`

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

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

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

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

Question 8. 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 attributes describe their shape, datatype, and the device.`

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

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

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

Question 9. 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 increase the precision of the model's predictions.`

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

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

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

Question 10. 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 shuffle the tokens in text embeddings`

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

Question 11. According to Lecture 9, what is the most effective way entity resolution differentiates between Jimmy Fallon and Jimmy Kimmel?

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

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

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

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

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

Question 12. 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 relies exclusively on rule-based methods, which are faster than statistical models.`

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

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

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

Question 13. 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) `Validation data is only used when there is not enough test data available.`

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

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

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

Question 14. 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) `Reduces the search space to identify the same entity.`

(d) `Compare records and identify matches.`

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

Question 15. 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 uses linear models while classification uses non-linear models`

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

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

(d) `Regression requires feature scaling while classification doesn’t`

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

Question 16. According to lecture 10, which Scikit-learn component is commonly used for converting text data into numerical vectors?

(a) `Vectorizer`

(b) `Pipeline`

(c) `Scaler`

(d) `Combining features into an colab notebook.`

(e) `Transformer`

Question 17. According to lecture 10, which of the following statements accurately describes the components of a confusion matrix in a binary classification task?

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

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

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

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

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

Question 18. According to Lecture 9: Entity Resolution, what is one key advantage of using factor graphs in entity resolution models?

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

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

(c) `They eliminate ambiguity entirely.`

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

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

Question 19. 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 and MAE always produce the same ranking of models, so either can be used interchangeably.`

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

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

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

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

Question 20. According to lecture 10, what does the R-squared score represent in a linear regression model ?

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

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

Question 21. 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 is only useful for linear models and not for non-linear models.`

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

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

Question 22. 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 separate data points into distinct classes using a linear function`

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

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

Question 23. According to lecture 10 (slide 26), what is the formula to calculate accuracy?

(a) `1 - SSres / SSTot.`

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

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

(d) `TP / TP + FP.`

(e) `TP / TP + FN.`

Question 24. According to lecture 9 on Entity Resolution, what is the primary challenge in resolving entity mentions across different text sources?

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

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

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

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

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

Question 25. According to lecture 10, which metric quantifies the proportion of correctly predicted positive instances among all instances predicted as positive?

(a) `Recall.`

(b) `Accuracy.`

(c) `F-score.`

(d) `Precision.`

(e) `Area Under the Curve (AUC).`

Question 26. 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 guarantees that every node is scanned multiple times to refine its cluster assignment.`

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

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

Question 27. 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 in different data formats.`

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

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

(e) `They are always identical.`

Question 28. According to lecture 8, which of the following best describes the primary benefit associated with token blocking?

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

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

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

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

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

Question 29. According to lecture 10, what is the purpose of the `make_regression` function in Scikit-learn API?

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

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

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

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

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

Question 30. According to lecture 8, which of the following is NOT true about blocking

(a) `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.`

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

(c) `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`

(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 31. 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 applies operations to all variables in an input matrix.`

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

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

Question 32. 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 assigns all mentions to a single entity immediately.`

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

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

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

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

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

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

(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 correctly predicts the positive class.`

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