cis6930sp25
March 7, 2025

Quiz 2 D

Question 1. 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) `Accuracy.`

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

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

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

Question 2. According to the lecture 11, what is the primary objective of Linear Regression in supervised learning?

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

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

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

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

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

Question 3. 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 takes features as input and returns a value for the target variable from a regression model.`

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

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

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

Question 4. 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) `Individual samples one at a time.`

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

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

Question 5. According to lecture 9, which blocking approach partitions the dataset based on a specific attribute and is noted for its intuitive implementation despite being sensitive to noise?

(a) `Standard blocking`

(b) `Hybrid blocking`

(c) `Token blocking`

(d) `Data profiling`

(e) `Sorted neighborhood`

Question 6. 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 positive linear but weak relationship with the actual values`

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

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

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

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

Question 7. 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 proportion of correctly classified positive instances among predicted positives .`

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

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

Question 8. According to lecture 10, what is the correct equation to solve for Precision?

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

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

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

(d) `TP / (TP + FN)`

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

Question 9. According to lecture 10, which of the following scenarios would maximizing Recall be more important than maximizing Precision?

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

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

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

(d) `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.`

(e) `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.`

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

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

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

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

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

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

Question 11. 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 only work for small datasets.`

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

(d) `They eliminate ambiguity entirely.`

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

Question 12. 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) `This approach never requires sorting the list of similar pairs prior to merging clusters.`

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

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

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

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

Question 14. 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's shape changes when a model changes the way it classifies only one outcome`

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

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

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

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

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

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

Question 16. 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 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 redundancy of using both coreference resolution and entity resolution in the same system.`

Question 17. 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 the algorithm requires all mentions to be merged in a single step.`

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

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

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

Question 18. 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 is simpler to implement and allows control over the mediator's behavior.`

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

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

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

Question 19. 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 shape of the curve does not change because the PR curve is independent of class distribution.`

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

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

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

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

Question 20. Which statement about the relationship between PR curves and ROC curves is MOST accurate?

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

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

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

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

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

Question 21. According to lecture 10, which of the following statements accurately describes the components of a confusion matrix in a binary classification task?

(a) `False Positives (FP) represent the number of positive instances 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 Negatives (FN) occur when positive instances are incorrectly classified as negative.`

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

Question 22. According to lecture 10, which of the following describes the purpose of a linear regression model?

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

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

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

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

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

Question 23. 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) `Precision increases, and Recall decreases.`

(e) `Both Precision and Recall decrease.`

Question 24. 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 order the data in a specific way (e.g. alphabetically or chronologically) such that the model can be trained more efficiently and faster.`

(b) `To vary the batch sizes`

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

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

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

Question 25. According to lecture 10, what does Precision measure?

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

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

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

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

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

Question 26. According to lecture 10, which metric quantifies the proportion of correctly predicted positive instances among all instances predicted as positive?

(a) `Precision.`

(b) `F-score.`

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

(d) `Accuracy.`

(e) `Recall.`

Question 27. 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 measures classification accuracy, making it suitable for both regression and classification tasks.`

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

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

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

Question 28. 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) `Tensors is a data structure analogous to a numpy array(1-D) or matrix(2-D), but can represent even higher dimensions.`

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

(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 29. 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 recall but may reduce precision due to redundant block membership.`

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

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

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

Question 30. 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 measures the average squared difference between the predicted and actual values.`

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

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

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

Question 31. 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) `Uses repel factors to distinguish between different entities, ensuring that mentions of Jimmy Fallon and Jimmy Kimmel do not get clustered together.`

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

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

Question 32. 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 is only useful for linear models and not for non-linear models.`

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

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

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

Question 33. 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 data formats.`

(b) `They are always identical.`

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

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

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