CS 506 PracticeFinal Exam

Lance Galletti

December 1, 2025

Name: _____

BUID: _____

Use the bubble sheet to mark your answers. Illegible answers will be treated as incorrect answers. There are 60 questions in total. You either know or you don't. All questions are worth 1pt. Incorrect answers are worth 0 pts. Good luck!

1 Correlation (2 questions)

1. The correlation coefficient of $y = \sin x$ for $x \in [0, 3\pi]$ is:

   A. 1

   B. 1/2

   C. 0

   D. -1/2

   E. -1

2. x, y, and z are features. The correlation between x and y is equal to that of y and z and equal to x and z. What is the minimum correlation it can be?

   A. -1

   B. $-1/2$

   C. 0

   D. 1/2

   E. 1

2 Clustering (7 questions)

3. Kmeans can, with poor initialization, never actually converge.

   A. True

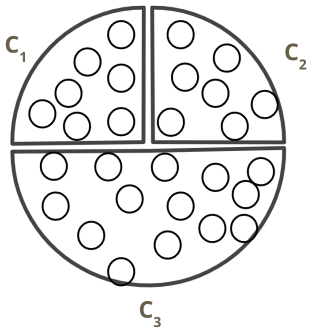
   B. False

4. Lloyd's algorithm always converges.

   A. True

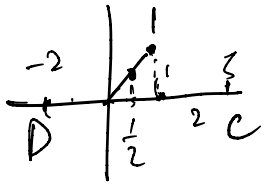
   B. False

5. Is the following a possible output from kmeans++ clustering?



A. Yes

B. No



6. Using euclidean distance and complete-link distance, what is the merging order of the following dataset? {A=(0,0), B=(1,1), C=(3,0), D=(0,-2)}

   A. AB → ABD → ABDC

   B. AB → ABC → ABCD

   C. AB → CD → ABCD

7. Keeping ϵ the same and increasing min_pts in DBScan will increase the number of noise points.

   A. True

   B. False

8. The Expectation Maximization Algorithm behind GMM is guaranteed to increase the likelihood of the data at each EM step.

   A. True

   B. False

9. How many parameters does GMM need to estimate?

   A. k

   B. 2*k

   C. 3*k

3 parameter

3 Singular Value Decomposition (6 questions)

10. As k increases, the frobenius distance between a matrix and its rank-k approximation decreases.

large k → smaller distance

   A. True

   B. False

11. In preparing data for dimensionality reduction, a researcher applies Singular Value Decomposition (SVD) directly to the data matrix X without mean-centering the features. They claim the resulting singular vectors will correctly capture the directions of maximum variance in the data.

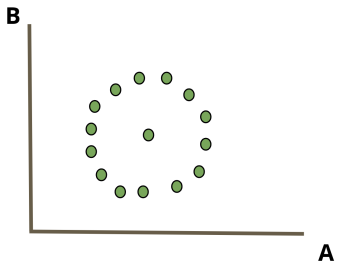
   A. True: SVD always finds the directions of greatest variance, centering is unnecessary.

without centering ✗ guarantee

   B. False: Without centering, SVD may align with the mean of the data rather than the true axes of variation.

12. The first principal component in SVD will align most with whichever feature has the largest variance.
    A. True
    B. False

13. What is the rank and dimension of the following dataset?



    A. rank 1 dimension 1
    B. rank 1 dimension 2
    C. rank 2 dimension 1
    D. rank 2 dimension 2

14. In SVD, U multiplied by Σ is a representation of our dataset in the dimensional space where each principal component is a basis vector.
    A. True
    B. False

15. Using SVD for feature selection is an example of
    A. Anomaly Detection
    B. Denoising
    C. Dimensionality Reduction

4 Classification (8 questions)

16. Overfitting is when
    A. Accuracy on the training set is high AND Accuracy on the testing set is low
    B. Accuracy on the training set is low AND Accuracy on the testing set is high

17. Underfitting is when
    A. Accuracy on the training set is high
    B. Accuracy on the training set is low

18. In KNN, when K=1, what is the training set accuracy?
    A. 100%
    B. 50%
    C. 0%
    D. Impossible to tell

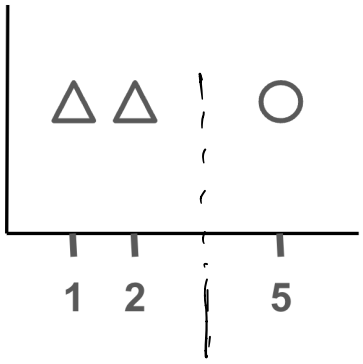
19. In KNN, the bigger the K the higher the chance of overfitting.
    A. True
    B. False

20. In KNN, the bigger the K the higher the chance of underfitting.
    A. True
    B. False

21. In KNN, if one of the features is income and its scale is much higher than other features, the nearest neighbors will almost always be the ones with the closest income regardless of other features.
    A. True
    B. False

22. Adjusting a model based on the testing set performance is a form of overfitting.
    A. True
    B. False

23. Where is the decision boundary of KNN for K=1 for the following dataset?



    A. 2.5
    B. 3
    C. 3.5
    D. 4

5 Decision Trees (5 questions)

24. What is the GINI of a node that only contains examples of a single class?

    A. 0

    B. .5

    C. 1

    D. Impossible to calculate with the information provided

25. Assuming binary classification, what is the GINI of a node with equal number of examples from each class?

    A. 0

    B. .5

    C. 1

    D. Impossible to calculate with the information provided

26. Splitting on an attribute with a bigger GINI will result in a worse overall decision tree

    A. True

    B. False

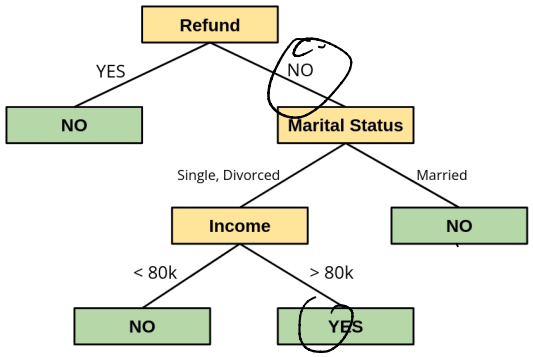
27. The GINI of a split is just the weighted average (weighted by the proportion of examples in each node) of the GINIs of each node created by the split.

    A. True

    B. False

28. Classify the following example using the decision tree below:

$$(\text{Refund} = \text{NO}, \text{Marital Status} = \text{MARRIED}, \text{Income} = 90\text{K})$$



    A. YES

    B. NO

    C. Impossible as we need to know the default class

6 Naive Bayes (4 questions)

29. Naive Bayes cannot be used if the features in the dataset are: [*total_price*, *unit_price*, *quantity*]

A. True

B. False

30. Naive Bayes assumes that all our features are independent

A. True

B. False

31. Estimate $P(\text{Attribute A} = Yes | \text{Class} = No)$ from the data.

Attribute A	Attribute B	Attribute C	Class
Yes	Single	High	No
No	Married	Mid	No
No	Single	Low	No
Yes	Married	High	No
No	Divorced	Mid	Yes
No	Married	Low	No
Yes	Divorced	High	No
No	Single	Mid	Yes
No	Married	Low	No
No	Single	Mid	Yes

A. 3/10

B. 3/7

C. 1

32. Using Naive Bayes and the same dataset as the previous question, classify (No, Divorced, Mid).

A. Yes

B. No

C. Impossible as we need to know the default class

7 Support Vector Machines (4 questions)

33. What is the width of the street for the following SVM: $3x_1 + 4x_2 + 1 = 0$

A. 1

B. 5

C. 2/5

D. $\sqrt{7}$

34. Multiplying an SVM by a constant $c < 1$ expands the SVM.

A. True

B. False

35. In SVM the variable Y that determines which class each data point belongs to takes values of either -1 or $+1$, not 0 or 1.

A. True

B. False

36. SVM uses all training data to create the decision boundary and street.

A. True

B. False

8 General Modeling (4 questions)

37. After splitting the data into a training and testing set and finding the model that gives the best results, it's best practice to merge the training and testing sets and retrain the model to maximize its potential.

A. True

B. False

38. Cross-validation is used to tune the parameters of the model and is only applied to the training set.

A. True

B. False

39. If the classes in the dataset are imbalanced, accuracy can be over inflated. In this case the following metric is preferred:

A. Precision

B. Recall

C. F1-score

40. Boosting and Bagging are both ensemble methods but

A. Bagging samples randomly with replacement while Boosting samples randomly but the weights of data points varies depending on how successfully these points were predicted

B. Bagging averages the predictions of all models while Boosting takes a weighted average.

C. Both of the above

D. None of the above

9 Regression (8 questions)

41. A mathematical function cannot have multiple Y (output) values for a single X (input) value but according to the assumptions of linear regression, our dataset can contain multiple different Y values for a fixed value of X.

A. True

B. False

42. According to the assumptions of linear regression, X is a random variable following a normal distribution.

A. True

B. False

43. According to linear regression, the function $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$ is a valid linear model that can describe Y.

A. True

B. False

44. In linear regression, finding $\hat{\beta}$ is an optimization problem with a unique solution.

A. True

B. False - if there are colinear features, there will be infinitely many solutions.

45. If X and Y are linearly dependent vectors then the fitted line from linear regression is exactly the first principal component of the matrix containing both X and Y.

A. True

B. False

46. A 99%confidence interval contains its target with 99% probability.

A. True

B. False

47. Out of 100 99% confidence intervals, 99 are expected to contain the true parameter value

A. True

B. False

10 Logistic Regression (9 questions)

48. In Binary Logistic Regression it is assumed that Y follows a Bernoulli Distribution with parameter $p = \sigma(X\beta)$.

    A. True

    B. False

49. In Binary Logistic Regression it is assumed that X follows a Logistic Distribution.

    A. True

    B. False

50. The Sigmoid Function is a Probability Density Function since it models probabilities

    A. True

    B. False

51. No matter which sigmoid-shaped function we use (think probit regression), the decision boundary will be linear just like standard logistic regression.

    A. True

    B. False

52. The Mean Squared Error can be used in Logistic Regression but since it creates a non-convex function because of the sigmoid function, it's usually avoided.

    A. True

    B. False

53. In logistic regression, points that are on the same line parallel to the decision boundary are assigned the same probability.

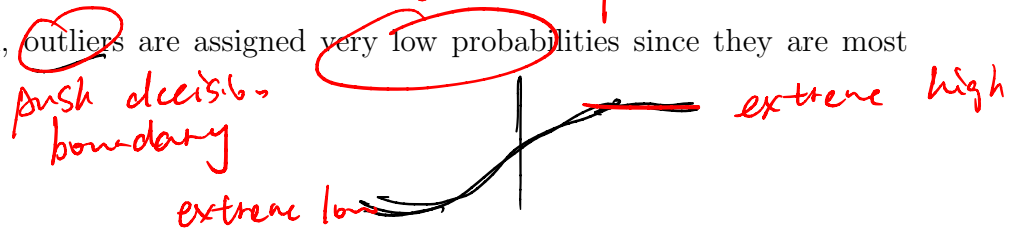
    A. True

    B. False

54. In logistic regression, outliers are assigned very low probabilities since they are most likely neither class

    A. True

    B. False



55. $P(Y|X)$ is a subject of interest in both Naive Bayes and Logistic Regression. The main difference is that Logistic Regression tries to estimate this probability directly while Naive Bayes does not.

    A. True

    B. False

11 Neural Networks (5 questions)

56. You train a fully-connected network with tanh activation functions on features and sigmoid activation functions in the output layer. The input features have large positive means. What will happen?

    A. It will work fine - neural networks can adapt to any features

    B. Most hidden features will be constant (either all 1 or all -1) on initialization take a very long time to change because the gradients are very small.

    C. The network will crash because finding the derivative of the sigmoid function is not possible.

57. In deep networks with sigmoid activations, neurons in later layers are more likely to become saturated because early layers map different inputs into similar directions, causing variance to shrink.

    A. True

    B. False

58. If two input features differ by a factor of 100 in scale, then backpropagation will automatically adjust the relevant weights during optimization, so feature scaling is not necessary.

    A. True

    B. False

59. A neural network with 1000 hidden neurons is more likely to overfit than a neural network with 100 hidden neurons.

    A. True

    B. False

60. Stochastic Gradient Descent could use a single example to update the weights instead of the entire batch.

    A. True

    B. False