Dept. of Mathematics and Statistics
King Fahd University of Petroleum & Minerals
AS491: Topics in AS & Fin. Math. II
Instructor: Dr. Ridwan A. Sanusi
Final Exam Term 252
Wednesday, May 20, 2026
7:00 PM - 9:30 PM

Name: _____ **ID#:** _____

Instructions:

1. No phones or any other smart devices. Any student caught during the exam will be considered under the cheating rules of the University.
2. Once the exam starts, nobody will be allowed to leave the exam room until the end of the exam.
3. Only materials provided by the instructor can be present on the table during the exam.
4. Do not spend too much time on any one question. If a question seems too difficult, leave it and go on.
5. While every attempt is made to avoid defective questions, sometimes they do occur. In the rare event that you believe a question is defective, the instructor cannot give you any guidance beyond these instructions.
6. Mobile calculators, I-pad, or communicable devices are disallowed. Use regular scientific calculators, financial calculators, or SOA-approved calculators only.
7. Answer all questions.
8. Formula sheet is provided on the next page.
9. Mark your answers clearly on the answer sheet.

Good Luck!

Formula Sheet

Concept	**Formula**
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
R^2	$1 - \frac{\text{RSS}}{\text{TSS}}$
F-statistic	$\frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$
LOOCV	$\frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i$
k-fold CV	$\frac{1}{k}\sum_{j=1}^{k}\text{MSE}_j$
LOOCV Shortcut (Linear)	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$
Bootstrap Inclusion Prob	$1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \approx 0.632$
C_p	$\frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$
AIC	$nlog(\text{RSS/n}) + 2d$
BIC	$nlog(\text{RSS/n}) + 2dlogn$
Adjusted R^2	$1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$
Ridge	$\text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$
Lasso	$\text{RSS} + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$
Gini Index (2-class)	$2\hat{p}(1-\hat{p})$
Entropy (2-class)	$-\hat{p}\log\hat{p} - (1-\hat{p})\log(1-\hat{p})$
K-Means Objective	$\sum_{k=1}^{K}\sum_{i\in C_k}\lVert x_i - \mu_k\rVert^2$
PC Score	$z_{i1} = \sum_{j=1}^{p}\phi_{j1}x_{ij}$
PVE	$\frac{\lambda_m}{\sum_{j=1}^{p}\lambda_j}$

1. A dataset has $n = 100$ observations and $p = 50$ predictors. You fit a ridge regression model with λ chosen by 10-fold CV. The resulting model has 50 non-zero coefficients. You then fit a lasso model with λ chosen by 10-fold CV. Which statement is MOST likely true?

   (A) Ridge will have more non-zero coefficients than lasso

   (B) Lasso will have more non-zero coefficients than ridge

   (C) Both will have exactly 50 non-zero coefficients

   (D) Ridge will have lower test MSE than lasso

   (E) Lasso will have lower test MSE than ridge

2. In K-means clustering, after convergence, you compute the total within-cluster sum of squares (WSS) and the between-cluster sum of squares (BSS). You know that TSS = WSS + BSS. If BSS/TSS = 0.85, what can you conclude?

   (A) The clustering explains 85% of the total variance

   (B) The clustering explains 15% of the total variance

   (C) The number of clusters K is 85

   (D) The clusters are poorly separated

   (E) The result is impossible because BSS cannot exceed WSS

3. You have a dataset with 3 predictors that are perfectly collinear: $X_3 = 2X_1 + 3X_2$. You fit a linear regression model using OLS. Which statement is TRUE?

   (A) The coefficients $\beta_1, \beta_2, \beta_3$ are uniquely determined

   (B) The predicted values $\hat{Y}$ are uniquely determined

   (C) The standard errors of the coefficients are valid

   (D) The R^2 cannot be computed

   (E) The F-statistic is undefined

4. In a random forest, a random sample of m predictors is selected from the full set of p predictors. Determine which of the following statements are true:

   I. The m predictors are sampled with replacement.

   II. The m predictors are sampled without replacement.

   III. A fresh sample of m predictors are selected for each split.

   IV. A fresh sample of m predictors are selected for each tree, but not for each split.

   (A) I, III

   (B) I, IV

   (C) II, III

   (D) II, IV

   (E) Fewer than two of the statements are true.

5. You perform PCA on a dataset with 5 variables after standardizing each to have mean 0 and variance 1. The eigenvalues are: $\lambda_1 = 2.5$, $\lambda_2 = 1.2$, $\lambda_3 = 0.8$, $\lambda_4 = 0.3$, $\lambda_5 = 0.2$. The cumulative proportion of variance explained by the first two PCs is closest to:

   (A) 0.50
   (B) 0.62
   (C) 0.74
   (D) 0.86
   (E) 0.94

6. In boosting, if the shrinkage parameter λ is set to 0.001 (very small) and the number of trees B is set to 10,000, what do you expect to happen?

   (A) The model will underfit severely
   (B) The model will overfit severely
   (C) The model will perform well because small λ prevents overfitting
   (D) The model will take a very long time to train but may perform well
   (E) The model will be equivalent to a single tree

7. Determine which of the following statements regarding the effect of omitting important variables from the regression model specification is NOT true.

   (A) It will result in underfitting
   (B) It will increase the width of prediction intervals
   (C) The coefficient estimates will remain unbiased
   (D) The estimate of the variance is inflated
   (E) The total sum of squares remains unchanged

8. You fit a lasso model and obtain coefficients: $\hat{\beta}_1 = 2.3$, $\hat{\beta}_2 = 0$, $\hat{\beta}_3 = 1.7$, $\hat{\beta}_4 = 0$, $\hat{\beta}_5 = 0.5$. You then increase the tuning parameter λ slightly. Which coefficient is MOST likely to become zero next?

   (A) $\hat{\beta}_1 = 2.3$
   (B) $\hat{\beta}_3 = 1.7$
   (C) $\hat{\beta}_5 = 0.5$
   (D) $\hat{\beta}_2 = 0$
   (E) Cannot be determined

9. In a classification tree, the root node has 200 observations (120 Class A, 80 Class B). A split creates left node (100 observations, 95 Class A, 5 Class B) and right node (100 observations, 25 Class A, 75 Class B). The reduction in Gini index is approximately:

   (A) 0.12 (B) 0.24 (C) 0.36 (D) 0.48 (E) 0.60

10. A dataset has 100 observations with six predictor variables, X1, X2, X3, X4, X5, and X6 and a response variable, Y. Three ordinary regression models have been run, with the following results:

Model Number	Variables Used	Residual Sum of Squares	Residual Standard Error
I	X1 only	183,663.30	43.2911
II	X1 and X2 only	8,826.47	9.5391
III	X1, X2, X3, X4, X5, and X6	8,319.59	9.4582

The Akaike Information Criterion (AIC) is to be used to select the best model from these three choices. Determine which of the following statements is true.

(A) Select Model I with an AIC value of 1,838.42
(B) Select Model I with an AIC value of 1,874.12
(C) Select Model II with an AIC value of 91.84
(D) Select Model II with an AIC value of 91.90
(E) Select Model III with an AIC value of 93.93

11. You are given a boosted regression tree with shrinkage parameter $\lambda = 0.08$. There is a single numeric predictor variable, x, and the target variable is y. Each tree is to have a single split. The first tree, built using all the training data has the following split:

$$\begin{aligned} &\text{If } x < 17.5: \quad \text{prediction} = 125.4 \\ &\text{If } x \geq 17.5: \quad \text{prediction} = 350.5 \end{aligned}$$

The following are three observations from the training set:

Observation	x	y
1	15.2	139
2	12.9	156
3	23.6	289

Calculate the updated predictions for these three observations

$$(\hat{f}(x_1), \hat{f}(x_2), \hat{f}(x_3))$$

(A) (–406.1, –389.1, –31.1)
(B) (–211.4, –194.4, 163.6)
(C) (10.0, 10.0, 28.0)
(D) (111.0, 128.0, 279.0)
(E) (129.0, 146.0, 261.0)

12. You perform PCA on a dataset and the first principal component loading vector is $\phi_1 = (0.5, 0.5, 0.5, 0.5)$. The second principal component loading vector must satisfy $\phi_2^T \phi_1 = 0$ and $\|\phi_2\| = 1$. How many possible loading vectors ϕ_2 (up to sign) satisfy these constraints?

    (A) 1

    (B) 2

    (C) 3

    (D) Infinite

    (E) None

13. In a random forest, the parameter m (number of predictors considered at each split) is set to p (all predictors). How does this compare to bagging?

    (A) The random forest becomes identical to bagging

    (B) The random forest becomes identical to a single tree

    (C) The random forest becomes identical to boosting

    (D) The random forest will overfit more than bagging

    (E) The random forest will underfit more than bagging

14. You have a dataset with 5 observations in 2D. The distance matrix (Euclidean) is given. Using complete linkage, the first fusion is between observations 1 and 2 at height 2. The second fusion is between observations 4 and 5 at height 3. The distance between clusters $\{1, 2\}$ and $\{4, 5\}$ is computed as 7. At what height do these two clusters fuse?

    (A) 2

    (B) 3

    (C) 5

    (D) 7

    (E) Cannot be determined

15. You fit a logistic regression model to predict heart disease. The coefficient for Age is 0.05 (p-value = 0.002). The coefficient for Cholesterol is 0.001 (p-value = 0.45). Which statement is CORRECT?

    (A) A one-year increase in age multiplies the odds of heart disease by $e^{0.05} \approx 1.05$

    (B) A one-year increase in age increases the probability of heart disease by 5 percentage points

    (C) Cholesterol is not associated with heart disease because its coefficient is small

    (D) The model predicts that older patients always have heart disease

    (E) The p-value for Age indicates the coefficient is practically significant

16. In K-means clustering with $K = 3$, you run the algorithm 20 times with different random initializations. The total within-cluster sum of squares values are: 145, 147, 146, 189, 148, 144, 190, 146, 147, 145, 191, 144, 146, 188, 147, 145, 192, 146, 189, 145. What is the most appropriate final clustering to report?

    (A) The clustering from the run with WSS = 144

    (B) The clustering from the run with WSS = 192

    (C) The average of all clusterings

    (D) The median of all clusterings

    (E) The clustering from the first run

17. A bootstrap sample of size $n = 1000$ is drawn with replacement. What is the expected number of observations that appear exactly once?

    (A) 368

    (B) 632

    (C) 1000

    (D) 500

    (E) 0

18. You have a dataset with $n = 200$ and $p = 5$. You fit a linear regression model and compute the AIC and BIC. Which statement is TRUE?

    (A) AIC and BIC will always select the same model

    (B) BIC will select a larger model than AIC

    (C) BIC will select a smaller model than AIC

    (D) AIC is unbiased for test error, BIC is not

    (E) Neither can be used for model selection

19. In a regression tree, the predicted value for a terminal node is the mean of the training responses in that node. If you change the prediction rule to the median instead of the mean, what happens to the training RSS?

    (A) It will always decrease

    (B) It will never decrease

    (C) It will never change

    (D) It may increase or decrease depending on the data

    (E) It will become zero

20. You have a dataset with two predictors that are highly correlated ($r = 0.95$). You fit three models: (1) OLS with both predictors, (2) ridge regression, (3) lasso. Which statement about the coefficient estimates is MOST likely true?

(A) OLS will give very large standard errors for both coefficients

(B) Ridge will make both coefficients exactly zero

(C) Lasso will make both coefficients exactly zero

(D) OLS will give smaller standard errors than ridge

(E) All three methods will give identical estimates

21. In hierarchical clustering with average linkage, the distance between clusters is defined as the average of all pairwise distances. Suppose clusters A and B have $|A| = 3$ and $|B| = 4$. How many pairwise distances are averaged?

(A) 7 (B) 12 (C) 21 (D) 3 (E) 4

22. You are given the following Lasso regularization path information: as λ increases from 0, the coefficient for X_1 becomes zero at $\lambda = 2$, for X_2 at $\lambda = 5$, for X_3 at $\lambda = 1$, and for X_4 at $\lambda = 10$. Which predictor is MOST important?

(A) X_1 (B) X_2 (C) X_3 (D) X_4 (E) Cannot be determined

23. In matrix completion, you initialize missing values with column means. After one iteration, you update the missing values using the first principal component. What happens to the column means in the next iteration?

(A) They remain exactly the same as the initial means

(B) They change because missing values have been updated

(C) They are reset to zero

(D) They become the principal component scores

(E) They converge to the true population means

24. You apply 2-means clustering to a set of five observations with two features. You are given the following initial cluster assignments:

Observation	$X1$	$X2$	Initial cluster
1	1	3	1
2	0	4	1
3	6	2	1
4	5	2	2
5	1	6	2

Calculate the total within-cluster variation of the initial cluster assignments, based on Euclidean distance measure.

(A) 32.0 (B) 70.3 (C) 77.3 (D) 118.3 (E) 141.0

25. You perform PCA on a dataset with $n = 50$ and $p = 30$ after standardizing. The first 10 principal components explain 95% of the variance. You then use these 10 PCs as predictors in a linear regression (PCR). How many parameters (including intercept) does the PCR model have?

(A) 10
(B) 11
(C) 30
(D) 31
(E) 50

26. In boosting, the residuals are updated as $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$. If $\lambda = 0.1$ and the current residual for an observation is $r_i = 10$, and the tree predicts $\hat{f}^b(x_i) = 8$, what is the updated residual?

(A) 2
(B) 8
(C) 9.2
(D) 10.8
(E) 18

27. A linear model has been fit to a dataset containing six predictor variables, F, G, H, I, J, and K. Determine which of the following statements regarding using Akaike information criterion (AIC) or Bayesian information criterion (BIC) to select an optimal set of predictor variables for this linear model is/are true.

I. AIC and BIC provide a direct estimate of the test error.
II. When choosing between the subsets F, G, H and I, J, K, AIC will always select the same subset as BIC.
III. For large sample sizes (n >7), the number of variables selected by BIC will be less than or equal to the number selected by AIC.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

28. You have a dataset with $n = 500$ and $p = 10$. You fit a random forest and a gradient boosted tree. The random forest has OOB error of 0.15. The boosted tree has 10-fold CV error of 0.14. Which statement is MOST appropriate?

(A) The boosted tree is better because its CV error is lower

(B) The random forest is better because OOB error is more reliable

(C) The comparison is invalid because different validation methods were used

(D) Both methods are equivalent within sampling variability

(E) The random forest will always outperform boosting

29. In K-means clustering, the objective function is $\sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$. If you increase K from 2 to 3, what happens to the objective value at convergence?

(A) It will always decrease

(B) It will always increase

(C) It may increase or decrease

(D) It will stay the same

(E) It will become zero

30. Determine which of the following statements about selecting the optimal number of clusters in K-means clustering is/are true.

I. K should be set equal to n, the number of observations.

II. Choose K such that the total within-cluster variation is minimized.

III. The determination of K is subjective and there does not exist one method to determine the optimal number of clusters.

(A) I only

(B) II only

(C) III only

(D) I, II and III

(E) The correct answer is not given by (A), (B), (C), or (D).

Q#	Answer	Explanation
1	A	Ridge never sets coefficients exactly to zero; lasso can. Ridge will have 50 non-zero, lasso likely fewer.
2	A	BSS/TSS is the proportion of variance explained by clustering (analogous to R^2).
3	B	Perfect collinearity means coefficients are not unique, but fitted values $\hat{Y}$ are uniquely determined.
4	C	II and III are true. Random forests require that the predictors be sampled without replacement. If not, the same predictor could be selected more than once and thus fewer than m predictors would be used. While it is possible to draw a fresh sample for each tree, that is not the definition of a random forest. (SRM Q74)
5	C	Total $= 2.5 + 1.2 + 0.8 + 0.3 + 0.2 = 5.0$. Cumulative $= (2.5 + 1.2)/5 = 3.7/5 = 0.74$.
6	D	Very small λ requires many trees (B) for good performance; training takes time but can work well.
7	C	The estimates will be biased. (SRM Q71)
8	C	The coefficient with smallest magnitude (0.5) is most likely to become zero next as λ increases.
9	B	Root Gini $= 2(0.6)(0.4) = 0.48$. Left Gini $= 2(0.95)(0.05) = 0.095$. Right Gini $= 2(0.25)(0.75) = 0.375$. Weighted avg $= 0.5(0.095) + 0.5(0.375) = 0.235$. Reduction $= 0.48 - 0.235 = 0.245 \approx 0.24$.
10	C	The AIC formula uses the estimated variance from the full model, sigma squared hat 89.45755. For each model, the AIC is: Model I: [183,663.30 + 2(1)(89.45755]/100 = 1,836.42. Model II: [8,826.47 + 2(2)(89.45755]/100 = 91.84. Model III: [8,319.59 + 2(6)(89.45755]/100 = 93.93. The model with the smallest AIC should be selected, which is Model II with AIC = 91.84. Note that the residual standard error did not have to be given. It can be calculated from the residual sum of squares, the sample size, and the number of predictors. (SRM Q70)
11	C	The initial fitted values are all zero They are updated by adding λ times the prediction from the tree. For the three observations the predictions are 125.4, 125.4, and 350.5. The updated predictions are 0 + 0.08(125.4) = 10.0, 0 + 0.08(125.4) = 10.0, and 0 + 0.08(350.5) = 28.0. (SRM Q69).
12	D	The constraint $\phi_2^T \phi_1 = 0$ defines a 3-dimensional subspace; infinite vectors satisfy it with unit norm.
13	A	When $m = p$, random forest considers all predictors at each split, same as bagging.
14	D	Complete linkage uses maximum distance; clusters fuse at height equal to their inter-cluster distance (7).
15	A	Logistic regression coefficient interpretation: odds multiply by e^{β} for one-unit increase.
16	A	Report the clustering with smallest WSS (best local optimum).
17	A	Probability of appearing exactly once $= n \cdot \frac{1}{n} \cdot (1 - \frac{1}{n})^{n-1} \approx e^{-1} \approx 0.368$, so expected $= 1000 \times 0.368 = 368$.
18	C	For $n > 7$, $\log n > 2$, so BIC penalizes more heavily and selects smaller models.
19	B	Median minimizes absolute deviation, mean minimizes squared error. Changing to median will increase RSS unless data are symmetric.
20	A	High correlation inflates standard errors in OLS; ridge reduces variance by shrinking coefficients.

Q#	Answer	Explanation
21	B	Number of pairwise distances = $\|A\| \times \|B\| = 3 \times 4 = 12$.
22	D	Predictor that remains non-zero for largest λ is most important. X_4 becomes zero at $\lambda = 10$ (largest).
23	B	Updated missing values change column means in next iteration.
24	C	The means for cluster 1 are $(1+0+6)/3 = 2.3333$ for X_1 and $(3+4+2)/3 = 3$ for X_2 and the variation is $(1-2.3333)^2 + (3-3)^2 + (0-2.3333)^2 + (4-3)^2 + (6-2.3333)^2 + (2-3)^2 = 22.6667$. The means for cluster 2 are $(5+1)/2 = 3$ for X_1 and $(2+6)/2 = 4$ for X_2 and the variation is 16. The total within-cluster variation (per equation (10.12) in ISLR) is $2(22.6667 + 16) = 77.33$. (SRM Q59)
25	B	PCR with $M = 10$ PCs has 10 slope coefficients + 1 intercept = 11 parameters.
26	C	$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) = 10 - 0.1 \times 8 = 10 - 0.8 = 9.2$.
27	D	I is false, AIC and BIC make an indirect estimate by adjusting the training error. II is true, for a fixed value of the number of predictors, the two provide the same ranking. III is true as for n >7, BIC provides a greater penalty for additional variables, and hence will select a number less than equal to that selected by AIC. (SRM Q61)
28	C	Different validation methods (OOB vs CV) give incomparable estimates. Need same validation method for fair comparison.
29	A	Adding more clusters can only decrease or keep same the objective (more flexibility always reduces within-cluster sum of squares).
30	C	I is false. Setting K = n will almost certainly overfit as there are unlikely to be that many true clusters. II is false. Within-cluster variation is minimized when K = n, which as noted above is unlikely to be optimal. III is true. There is no exact method for determining the optimal value of K. (SRM Q60)