STA 314H1S: Midterm Exam

Summer, 2023

Instructor: Ziang Zhang

Time allowed: 3 hours

Aid allowed: One non-programmable calculator

Instructions

- Students are expected to complete *all* the questions within the space provided. If the extra page is used, please indicate clearly which question is being answered.
- Write your answers clearly and show your detailed steps. If the examiner cannot read an answer or follow your steps, it will be marked as incorrect.
- You have to write your answers on the QR-coded pages. The non-QR coded pages can be used as scratch papers.
- For multiple choice questions, remember to fill your answers on the bubble sheet. You are suggested to fill the answers using pencils, so you can still change them.
- The total mark in this test is 100. Good luck!

Multiple Choices (20 pts): For each of these questions, select *the best* choice.

1. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, which of the following statements is *incorrect*?

    (a) If $\boldsymbol{A}$ is a positive semi-definite (PSD) matrix and c is a scalar, $c\boldsymbol{A}$ may not be a PSD matrix.

    (b) If $\boldsymbol{A}$ is a PSD matrix and $\boldsymbol{B}$ is a positive definite (PD) matrix, $\boldsymbol{A}+\boldsymbol{B}$ must be a PD matrix.

    (c) If $\boldsymbol{A}$ is a symmetric matrix, then $\boldsymbol{A}$ has an eigen-decomposition.

    (d) If $tr(\boldsymbol{A}) \geq 0$, $\boldsymbol{A}$ must be a PSD matrix.

2. Assume $\boldsymbol{x} \in \mathbb{R}^p$ is a (multivariate) normal vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, which of the following statements is *incorrect*?

    (a) If $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{a}^T\boldsymbol{x}$ has a variance of $\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$.

    (b) The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may not specify the entire distribution of $\boldsymbol{x}$.

    (c) $\boldsymbol{\Sigma}$ must be a symmetric matrix with size $p \times p$.

    (d) If $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathbf{A}\boldsymbol{x} \in \mathbb{R}^m$ is also a normal vector.

3. Assume $\boldsymbol{x} \in \mathbb{R}^{p_1}, \boldsymbol{y} \in \mathbb{R}^{p_2}$ are random vectors with means respectively being $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ and covariance matrices respectively being $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$, which of the following statements is *incorrect*?

    (a) If $p_1 = p_2$, $\boldsymbol{x}+\boldsymbol{y}$ has mean $\boldsymbol{\mu}_x + \boldsymbol{\mu}_y$.

    (b) $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ must be a PSD matrix.

    (c) If $\boldsymbol{x}$ and $\boldsymbol{y}$ are both normal, $\mathbf{x}+\mathbf{y}$ may not be normal.

    (d) If $\boldsymbol{x}, \boldsymbol{y}$ are independent, $\mathrm{Cov}(\boldsymbol{x}-\boldsymbol{y}) = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y$.

4. Consider the following statements related to the training, validation, and testing sets in the context of model fitting and selection. Which of these is *incorrect*?

    (a) It is generally advisable to use the testing set multiple times to tweak the model's parameters and hyperparameters for better performance.

    (b) The training set is used to fit the parameters of the model.

    (c) The validation set is used to tune hyperparameters and decide on the complexity of the model.

    (d) The testing set is used to provide an unbiased estimate of the model's performance on unseen data.

5. Which of the following is an assumption used for the testing MSE decomposition in the lecture?

   (a) The prediction $\hat{y}$ is constructed from a linear regression model.

   (b) The testing outcome and the prediction $\hat{y}$ are independent.

   (c) The sample size of the training data must be larger than the number of parameters in the model.

   (d) The feature x has been centered to have zero mean.

6. Consider the K-nearest neighbors (KNN) prediction method and its (test) MSE decomposition. Which of the following statements is *correct* regarding the choice of K?

   (a) Increasing K will always decrease both the bias and variance components of the MSE.

   (b) Decreasing K will always increase both the bias and variance components of the MSE.

   (c) Decreasing K will decrease the Bayes error of the KNN model.

   (d) Increasing K will decrease the variance component but increase the bias component of the MSE.

7. Which of the following statement best compares the linear regression method with the KNN regression method?

   (a) Linear regression is a non-parametric method, while KNN is a parametric method.

   (b) If you have more features than the number of training observations, the KNN method cannot produce a prediction.

   (c) If there is a one feature $x \in \mathbb{R}$, linear regression produces $\hat{y}(x)$ as a wiggly, discontinuous function, while KNN can only create continuous $\hat{y}(x)$.

   (d) Linear regression assumes a specific functional form for the regression function, while KNN does not make such strong assumptions and can adapt to various pattern in the data.

8. Assume you are using p features to predict an outcome variable y through the linear regression method. The training outcome is denoted as $\mathbf{y} \in \mathbb{R}^n$ and the design matrix is $\mathbf{X} \in \mathbb{R}^{n \times p}$, then:

   (a) The MLE $\hat{\boldsymbol{\beta}}$ is undefined when $p > n$.

   (b) The MLE $\hat{\boldsymbol{\beta}}$ is undefined when some features are perfectly correlated.

   (c) Regardless what are the values of $\boldsymbol{y}$, the MLE $\hat{\boldsymbol{\beta}}$ is well defined as long as $\mathbf{X}^T\mathbf{X}$ is a PD matrix.

   (d) All the statements are correct.

9. Suppose you construct a prediction $\hat{y}(x)$ from the training set $\{x_i, y_i\}_{i=1}^n$, select the *correct* statement from below.

   (a) The testing MSE is monotonic function as the complexity of the model.

   (b) The Bayes error could go up if the number of parameters in the model increases.

   (c) The training MSE can reflect the bias of $\hat{y}(x)$

   (d) The testing MSE can be computed as the sum of squared bias and variance of $\hat{y}(x)$.

10. Which of the following fact is *incorrect* about the model regularization?

    (a) For certain models, regularization allows a model to have a larger number of features than the number of observations.

    (b) Regularization can help prevent overfitting by adding a penalty term to the loss function, discouraging complex models.

    (c) Adding regularization typically reduces the performance of models on the training data.

    (d) Regularization can only be used in the context of linear regression.

11. Which of the following is an algorithm to *train* the Elastic Net regression?

    (a) Fisher Scoring

    (b) Leave-one-out CV

    (c) Coordinate Descent

    (d) Gradient Descent

12. Which of the following is an algorithm to *tune* the hyper-parameter λ in a LASSO regression?

    (a) K-fold CV

    (b) Fisher Scoring

    (c) Coordinate Descent

    (d) Gradient Descent

13. Assume $\lambda > 0$ and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3$, which of the following will not be a penalty on the norm of $\boldsymbol{\beta}$?

    (a) $\lambda\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}$

    (b) $\lambda|\beta_1 + \beta_2 + \beta_3|$

    (c) $\lambda(|\beta_1^2| + |\beta_2| + \beta_3|)$

    (d) $\lambda\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}$ for some PD matrix $\mathbf{A}$.

14. Assume $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, which of the following statements about PCA is *incorrect*?

    (a) If the data is centered, the (sample) PC directions can be obtained as the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

    (b) If the rows of $\mathbf{X}$ are *i.i.d* normal random vectors, all the PC scores must be normal as well.

    (c) If the data is centered, the variance of each (sample) PC score will be the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

    (d) The maximum number of PCs that can be calculated is p, the number of variables or columns in the design matrix.

15. Which of the following statements about Logistic Regression is incorrect?

    (a) Logistic regression can be used to construct a probabilistic classifier.

    (b) Logistic regression can produce a probability estimate for $\mathrm{P}(Y = 1|X = x_0)$ at a feature value x_0.

    (c) Logistic regression assumes the probability $\mathrm{P}(Y = 1|X)$ being a linear function.

    (d) The parameters in logistic regression can be estimated through MLE.

16. Which of the following does not produce a generative classifier?

    (a) Logistic regression

    (b) Linear discriminant analysis

    (c) Quadratic discriminant analysis

    (d) Naive Bayes

17. Which of the following statements about Naive Bayes classifier is *incorrect*?

    (a) Naive Bayes classifier assumes that all predictors are independent of each other, given the class.

    (b) The Naive Bayes assumption in practice can often reduce the number of parameters in the feature distribution $\mathrm{P}(\mathbf{x})$ significantly.

    (c) Naive Bayes classifier models the joint distribution of features and class labels, and makes predictions by applying Bayes' theorem to derive the conditional probability of a class given the features.

    (d) Naive Bayes classifier can only be used for binary classification problems.

18. Which of the following statements is *incorrect* about the LDA and QDA classifications?

    (a) In LDA, the decision boundary between two classes is linear in the predictors, while in QDA, the decision boundary is quadratic.

    (b) LDA assumes that the classes have the same covariance matrix, while QDA allows for each class to have its own covariance matrix.

    (c) If the assumption of a common covariance matrix is violated, QDA will likely outperform LDA because of its smaller variance.

    (d) The computation cost of LDA is typically lower than QDA.

19. You are developing a new diagnostic test for COVID-19. Which of the following statements is *incorrect*?

    (a) If the sensitivity of the test is high, it means the test is good at identifying patients who have COVID.

    (b) If the specificity of the test is high, it means the test is good at identifying patients who do not have COVID.

    (c) If the prevalence of COVID in the population is very low, even a test with high sensitivity and specificity could still have a very high overall error rate.

    (d) In a population with high COVID prevalence, a positive test result is more likely to be correct.

20. Which of the following statement is wrong about AUC and ROC?

   (a) The value of AUC can be computed from the ROC.

   (b) The ROC can be constructed from the value of AUC.

   (c) If your feature is not related to your outcome, the AUC of your classifier still likely be positive.

   (d) If your classifier does a perfect job at predicting the outcome, its AUC value should be 1.

1. (15 pts) For each of the following statements, either prove it is true using the basic properties and definitions covered in classes, or provide a counterexample.

   (a) (5 pts) If $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix with eigenvalues $\{\lambda_i\}_{i=1}^n$, then its trace can be computed as $tr(A) = \sum_{i=1}^n \lambda_i$.

   (b) (5 pts) If $A \in \mathbb{R}^{n \times n}$ is a positive semi-definite (PSD) matrix, then $A + I_n$ is a invertible matrix, where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix.

   (c) (5 pts) Assume $\mathbf{x} \in \mathbb{R}^p$ is a random vector with zero mean and identity covariance matrix I_p. Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be two orthogonal vectors in $\mathbb{R}^p$, then $\mathbf{a}_1^T\mathbf{x}$ and $\mathbf{a}_2^T\mathbf{x}$ are uncorrelated.

2. (15 pts) (*Regularized Regression*) Suppose you collect the following training data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ with sample size n. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the design matrix, with its ith row being $\boldsymbol{x}_i^T$, and $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ be the vector of training outcomes. You can assume that the model does not have the intercept parameter β_0.

   (a) (3 pts) Let $\lambda \geq 0$ be the penalty parameter. Write the regularized loss function for Ridge regression. Also, write out the Ridge regression estimate $\hat{\boldsymbol{\beta}}_{reg}$. Discuss whether this estimate will always be well defined even if $p \geq n$.

   (b) (2 pts) Rewrite the regularized loss function and the Ridge regression estimate from the previous part in terms of an un-regularized OLS loss function and an OLS estimate for an augmented dataset. Clearly define and describe the *augmented dataset* in this context, including the dimensions of them.

(c) (5 pts) Consider the following three R outputs of estimated beta coefficients:

```
Output A:
Intercept    Beta1    Beta2    Beta3   ...   Beta9    Beta10
3.3          1.1      -1.5      1.4    ...   -1.2     0.8
Output B:
Intercept    Beta1    Beta2    Beta3   ...   Beta9    Beta10
3.3          1.4      -2.5      1.6    ...   -1.3     0.8
Output C:
Intercept    Beta1    Beta2    Beta3   ...   Beta9    Beta10
3.3          0.0      -1.7      1.6    ...   0.0      0.0
```

Which of these outputs corresponds to each of the following methods? Briefly explain why.

(i) LASSO with $\lambda = 1$

(ii) Ridge with $\lambda = 1$

(iii) Ridge with $\lambda = 5$

Provide your answers along with justifications.

(d) (5 pts) As the regularization parameter λ in ridge regression varies from 0 to ∞, describe how each of the following quantities is expected to change, and provide a brief justification for each:

- Squared Bias
- Variance
- Bayes Error
- Testing MSE
- Training MSE

3. (15 pts) (*PCA*) Assume you have collected n pairs of *iid* feature vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ in $\mathbb{R}^p$. They are stacked together into one design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. You want to develop a predictive model using the PC regression method.

   (a) (5 pts) Below is a chunk of R code intended to compute the first k PC directions of the design matrix $\mathbf{X}$. However, there is a bug in this code. Identify the bug and explain how you could fix it.

```
performPCA <- function(X, k) {
  # Perform eigen decomposition
  eigen_decomp <- eigen(t(X) %*% X)

  # Return first k PC directions
  return(eigen_decomp$vectors[, 1:k])
}
performPCA(X, k)
```

(b) (5 pts) Suppose your colleague constructs a PC regression model using the first k_1 PCs, where k_1 is selected based on the scree plot. You, on the other hand, plan to build a model using the first k_2 PCs, where k_2 is determined through K-fold cross-validation (CV).

Outline the procedure or steps you would follow to select this optimal number of PCs, k_2, using K-fold CV. If you find it helpful, you may write your answer in the form of pseudo-code.

(c) (3 pts) Continuing from the previous question, assume $k_1 > k_2$. Which model—your model or your colleague's—do you expect to have better predictive performance on the test set? And which one do you anticipate to have a better performance on the training set? Provide a brief explanation for both of your answers.

(d) (2 pts) In which of the following scenarios do you think PC regression will be more useful: when features are highly correlated, or when features are largely independent of each other? Briefly explain your reasoning.

4. (15 pts) *(Generative Method)* Assume you want to derive a generative classifier for an outcome variable y that takes two possible levels $k = 0, 1$. Conditional on $y = k$, the feature variable $x \in \mathbb{R}$ is normally distributed with mean μ_k and variance σ^2, with density

$$f(x|y = k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma^2}\right).$$

(a) (5 pts) Let $\pi_k = P(y = k)$, show that the discriminant function is

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

(b) (5 pts) Assume $\pi_0 = 0.5$ and $\mu_0 < \mu_1$, show that the Bayes decision boundary is defined by the point $x^* = \frac{\mu_0+\mu_1}{2}$. When will an observation be classified as level 1 using this Bayes classifier?

(c) (5 pts) Continue from part (b), show that the same Bayes classifier can be obtained from a logistic regression model:

$$\log\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = \beta_0 + \beta_1 x,$$

and write β_0 and β_0 as functions of μ_0, μ_1 and σ.

5. (20 pts) (*Discriminative Method*) Assume the following logistic regression model is used to construct a classifier for $Y = 0, 1$ using a feature $x \in \mathbb{R}$:

$$\log \left[\frac{P(Y = 1|x)}{1 - P(Y = 1|x)} \right] = \beta_0 + \beta_1 x.$$

Assume you have a training set $\{y_i, x_i\}_{i=1}^n$ with n independent pairs of data.

(a) (5 points) Write out the likelihood and hence the log-likelihood functions for the given training set. Explain whether the MLE of this likelihood is always unique. If yes, provide a reason. If not, specify the condition under which it is not unique.

(b) (5 points) Druing the lecture, we covered an iterative algorithm that can be used to numerically find the MLE of the likelihood function in (a). Describe how will you implement that algorithm to find the MLE $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$. If you find it helpful, you may write your answer in the form of pseudo-code.

(c) (4 points) Given the MLE $\hat{\boldsymbol{\beta}} = (0.5, -0.1)$, the following R code is used to construct a Bayes classifier at a new feature value $x_0 = 1.2$.

```
beta_hat <- c(0.5,-0.1) # The MLE beta vector
x0 <- 1.2 # The new feature value
predicted_prob <- exp(beta_hat[1] + beta_hat[2] * x0)
predicted_y <- ifelse(predicted_prob > 0.5, 1, 0)
```

There exists a bug in the code that invalidates the classifier. Explain what is that bug and how will you fix it. Also describe what effect will this bug cause for your classifier.
(*Note:* The `ifelse` function in R checks the condition `predicted_prob > 0.5`. If it's TRUE, then `1` is returned; otherwise, `0` is returned.)

(d) (6 points) Suppose you have successfully built a Bayes classifier, and you have run your classifier on an independent test data. You get the following contingency table for your Bayes classifier:

	Predicted: 0	Predicted: 1	Total
Actual: 0	30	10	40
Actual: 1	20	40	60
Total	50	50	100

Compute the following quantities of your classifier:

i. Overall Error rate.

ii. Specificity.

iii. Sensitivity.

Describe how these quantities would likely change if you change your classification threshold so $\hat{y} = 1$ if and only if the estimated $P(y = 1|x)$ is larger than 0.9.

Extra page:

Extra page:

Extra page:

Extra page:

Extra page: