STA 314H1S: Midterm Exam (Practice)

Summer, 2023

Instructor: Ziang Zhang

Student Name: _____

Student ID: _____

Instructions

- Time allowed: 3 hours; Total points: 100.
- One non-programmable calculator is allowed.
- Students are expected to complete *all* the questions within the space provided. Extra space is provided at the end of the exam. If the extra space 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.
- Good luck!

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

1. Let $A, B, C \in \mathbb{R}^{n\times n}$, which of the following statement is *correct*?

   (a) $\mathrm{tr}(A)$ is only defined if A is a symmetric matrix.

   (b) $\mathrm{tr}(AC) = \mathrm{tr}(CA)$ always hold.

   (c) $\mathrm{tr}(AB) = \mathrm{tr}(A)\mathrm{tr}(B)$ always hold.

   (d) $\mathrm{tr}(ABC) = \mathrm{tr}(BAC)$ always hold.

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

   (a) $|A| = 0$ implies A is not invertible.

   (b) $\mathrm{tr}(A) = 0$ implies A is not invertible.

   (c) If both A, B are invertible, $A + B$ may be non-invertible.

   (d) If both A, B are invertible, AB must be invertible.

3. Assume $\boldsymbol{x} \in \mathbb{R}^p$ is a random vector with mean μ and covariance matrix Σ, which of the following statement is *incorrect*?

   (a) Σ must be a positive semi-definite matrix.

   (b) The parameters μ and Σ are enough to specify the entire distribution of $\boldsymbol{x}$.

   (c) Σ must be a symmetric matrix with size $p \times p$.

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

4. Assume $\boldsymbol{x} \in \mathbb{R}^{p_1}$ has mean μ_x and covariance matrix Σ_x, and $\boldsymbol{y} \in \mathbb{R}^{p_2}$ has mean μ_y and covariance matrix Σ_y, which of the following statement is *correct*?

   (a) $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ is a positive semi-definite matrix.

   (b) If both $\boldsymbol{x}$ and $\boldsymbol{y}$ are normal, the stacked vector $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p_1+p_2}$ is also normal.

   (c) If $\boldsymbol{x}$ is independent of $\boldsymbol{y}$, all the entries of $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ must be zero.

   (d) $\boldsymbol{x}$ is independent of $\boldsymbol{y}$ if and only if all the entries of $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ are zeros.

5. Assume $\boldsymbol{x} \in \mathbb{R}^p$ has mean μ_x and covariance matrix Σ_x, which of the following statement is *incorrect*?

   (a) $\mathbb{E}(\text{tr}(\boldsymbol{x}\boldsymbol{x}^T)) = \text{tr}(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T))$.

   (b) $\boldsymbol{x}^T\boldsymbol{x}$ is a non-negative random variable in $\mathbb{R}$.

   (c) For any $A \in \mathbb{R}^{p\times p}$ that is positive definite, $\boldsymbol{x}^T A\boldsymbol{x}$ is a positive random variable in $\mathbb{R}$.

   (d) $(\boldsymbol{x} - \mu_x)^T\Sigma_x^{-1}(\boldsymbol{x} - \mu_x) \sim X^2(p)$.

6. Consider a random vector (y, x) that follows (multivariate) normal distribution where x is the feature and y is the outcome variable to be predicted. You can assume that the feature x has been observed, and hence is a fixed value. The target is to derive the best deterministic prediction $\hat{y}(x)$ in terms of minimizing the MSE (i.e. $\hat{y}(x)$ is not a random variable). Select the *best* statement from below.

   (a) $\hat{y}(x)$ can never be $\mathbb{E}(y)$, hence the information of x is not used.

   (b) The best deterministic prediction in terms of MSE is not necessarily unique.

   (c) $\hat{y}(x)$ must be a certain linear regression model $\hat{y}(x) = \beta_0 + \beta_1 x$.

   (d) In order to derive $\hat{y}(x)$, it is important to know $\text{Var}(y|x)$.

7. Assume you want to build a linear regression model to predict the final exam grade of an STA314 student based on his/her grades from the five pre-requisites of this course (i.e. $p = 5$). Suppose you can access the final grades of $n = 100$ students in the last year's STA314 class. Select the *most accurate* statement from below.

   (a) You should fit models with different features using the $n = 100$ students, and select the model in terms of the smallest residual standard deviation.

   (b) The regression coefficients fitted through MLE may not be the same as the ones fitted through OLS.

   (c) You don't need to worry too much about the correlations within the p features, since $p < n$ ensures the existence of MLE in linear regression.

   (d) As the difficulty/relevance of the courses in the previous year may be different from this year, some predictive features in the previous year may be no longer predictive now.

8. Select the *correct* statement from below.

   (a) Compared to the nonparametric method, the parametric method benefits from faster computation and easier interpretability.

   (b) The KNN regression is a nonparametric method, with its bias-variance tradeoff controlled by the choice of K.

   (c) The linear regression method is a parametric method, which assumes the outcome variable is linearly affected by the regression coefficients $\boldsymbol{\beta}$.

   (d) All the statements are correct.

9. Suppose we want to predict a random variable y with a prediction $\hat{y}(x)$, select the *correct* statement from below.

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

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

   (c) The MSE of $\hat{y}(x)$ should be larger than the sum of its squared bias and its variance.

   (d) The MSE of $\hat{y}(x)$ should be larger than the MAE of $\hat{y}(x)$.

10. Select the *correct* statement from below.

    (a) LASSO is a type of regularization method that applies the L2 norm penalty to the objective function.

    (b) LASSO is a type of regularization method that produces sparsity in the estimated regression coefficient, and hence is guaranteed to perform better than the unregularized linear regression model fitted through OLS.

    (c) For linear regression, the Ridge regularization method always provides a unique solution even if $p \geq n$.

    (d) When the original model is linear regression, neither LASSO nor Ridge permits a closed-form solution.

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

    (a) Coordinate Descent

    (b) Gradient Descent

    (c) Fisher Scoring

    (d) Leave-one-out CV

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

    (a) Coordinate Descent

    (b) Gradient Descent

    (c) Fisher Scoring

    (d) Leave-one-out CV

13. Which of the following is the main advantage of the elastic net over LASSO?

    (a) The elastic net can produce sparser estimates of the regression coefficients.

    (b) The computation of the elastic net method is significantly easier than LASSO.

    (c) The elastic net has a more stable behavior than LASSO for correlated covariates.

    (d) The prediction yielded by the elastic net is more accurate.

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

    (a) 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 rows of $\mathbf{X}$ are $i.i.d$ normal random vectors, all the PC scores must be independent.

    (d) The total variance explained by the first p PCs is likely larger than the total variance of the original features.

15. Assume we are currently using p features to predict an outcome variable y. Which of the following statements about PC regression is *correct*?

    (a) Compared to the original regression, regressing y on the first $k < p$ principal components of the features likely decreases the training MSE.

    (b) Compared to the original regression, regressing y on the first $k < p$ principal components of the features likely decreases the testing MSE.

    (c) Compared to the original regression, regressing y on the p principal components of the features likely performs better in terms of prediction accuracy.

    (d) PC regression could be a valuable tool when $p \approx n$, where n denotes the sample size.

16. Which of the following method produces a generative classifier?

    (a) Linear regression

    (b) Logistic regression

    (c) K-Nearest Neighborhood

    (d) Naive Bayes

17. Which of the following is a disadvantage of using logistic regression for classification?

    (a) Logistic regression cannot guarantee to produce a probability estimate in $[0, 1]$.

    (b) Logistic regression cannot be used together with shrinkage methods such as LASSO.

    (c) Logistic regression does not have a unique MLE when y's are perfectly separated.

    (d) Logistic regression has no limitation; It is the best approach in the world.

18. Suppose you are constructing a *generative* classifier using a very large number of features $x_1, ..., x_p$, where some of them are discrete and some of them are continuous. The sample size n is also huge which makes the computation efficiency a crucial factor to consider. Which of the following approach is more suitable to start with in this case?

    (a) Logistic regression

    (b) Linear discriminative analysis

    (c) Quadratic discriminative analysis

    (d) Naive Bayes method

19. After you construct your initial classifier, you evaluate its misclassification rate using a validation dataset. You noticed that although the misclassification rate is small, a majority of the outcomes in the reference group are misclassified (i.e. $y_i = 0$ but $\hat{y}_i = 1$). Which of the following statement must be correct?

    (a) There must be an extreme imbalance between the number of $y = 0$ and the number of $y = 1$ in the validation set.

    (b) Your classifier has a high sensitivity but a low specificity.

    (c) To mitigate this problem, you could consider increasing the classification threshold for $P(y_i = 1|\boldsymbol{x})$ for $\hat{y}_i = 1$.

    (d) All the statements above are correct.

20. You decide to evaluate the performance of your classifier under a range of classification thresholds for $P(y_i = 1|\boldsymbol{x})$, and hence you compute its ROC and AUC for a validation set:

    (a) If your features are totally unrelated to your target outcome, the ROC is likely below the line $y = x$ and your AUC should be close to 0.

    (b) If you have constructed a very accurate classifier, but you accidentally flip the labels of your prediction (i.e. 1 to 0 and 0 to 1), then your AUC should be close to 0.5.

    (c) If you construct a random classifier $\hat{y}$, which is a Bernoulli random variable independent of the features, with $P(\hat{y} = 1)$ being the classification threshold, then your AUC should be close to 0.5.

    (d) All the statements above are correct.

1. (15 pts) For each of the following statements in linear algebra, either prove it is true using the basic properties and definitions discussed in class, or provide a counter-example showing it is false.

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

   (b) (5 pts) If $A \in \mathbb{R}^{n \times n}$ is an invertible matrix, $\text{tr}(A)$ cannot be zero.

   (c) (5 pts) If $A \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix, A^p will also be a positive semi-definite matrix for any positive integer p.

2. (20 pts) (*Regularized Regression*) Suppose the data is generated from the following linear regression model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Denote the training data as $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ where n is the sample size. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the design matrix, with its ith row being $\boldsymbol{x}_i^T$.

(a) (3 pts) Let $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$. What is the conditional distribution of $\mathbf{y}$ given all the features $\mathbf{X}$? Can you also conclude the marginal distribution of $\mathbf{y}$ based on the information above?

(b) (5 pts) Assume $\mathbf{X}^T\mathbf{X}$ is invertible, derive the expression of the MLE $\hat{\boldsymbol{\beta}}_{ML}$ for $\boldsymbol{\beta}$. Recall for a random variable $\mathbf{z} \sim N(\mu, \boldsymbol{\Sigma})$, its density is:

$$f(\mathbf{z}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right),$$

(c) (5 pts) Unfortunately you cannot collect too much data due to the tight budget, hence $n < p$. Show that $\hat{\boldsymbol{\beta}}_{ML}$ above is not uniquely defined (*hint: you just need to show* $\mathbf{X}^T\mathbf{X}$ *is not invertible*).

(d) (7 pts) To solve the above problem, you decide to use a Ridge regression to obtain the estimate $\hat{\boldsymbol{\beta}}_{Reg}$. Derive the formula of $\hat{\boldsymbol{\beta}}_{Reg}$ and explain why it is well-defined even if $\mathbf{X}^T\mathbf{X}$ is not invertible (*hint:* when deriving $\hat{\boldsymbol{\beta}}_{Reg}$, you can take the original loss function to be either the negative log-likelihood or the RSS).

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. For simplicity, assume the feature vectors are already centered to have zero means.

   (a) (3 pts) Your collaborator told you that if you only want to keep k terms in your final model, performing a best-subset selection with size k on the original features or a PC regression using the first k PCs will be equivalent since the first k PCs are designed to optimize the variance explained. Do you agree with your collaborator? If not, explain using no more than two sentences why your collaborator is wrong.

   (b) (6 pts) Derive the expressions of the first (sample) PC direction $\mathbf{v}_1$. Given the kth PC direction $\mathbf{v}_k$, how will you compute the kth PC score z_{ik} of the observation $\mathbf{x}_i$?

(c) (6 pts) Now assume the covariance matrix $\mathbf{\Sigma_X}$ is known, and you have computed (population) PC directions $\{\mathbf{v}_k\}_{k=1}^p$ using the true covariance. Prove that the two resulting (population) PC scores z_{ik} and z_{ij} (where $k \neq j$) are uncorrelated. Do we need any additional assumptions to claim that they are independent?

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 vector $\mathbf{x} \in \mathbb{R}^2$ is normally distributed with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, with density

$$f(\mathbf{x}|y = k) = \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

   (a) (5 pts) Derive the quadratic discriminant function $\delta_k(\mathbf{x})$, and show maximizing δ_k is equivalent to maximizing the probability $P(y = k|\mathbf{x})$.

(b) (10 pts) Assume you have observed $n = 10$ data points, with 4 observations in the 0 class, 6 observations in the 1 class.

The estimated means are $\bar{\boldsymbol{\mu}}_0 = [-3/2, 3/2]^T, \bar{\boldsymbol{\mu}}_1 = [1/2, 1]^T$.

The estimated covariances are $\hat{\boldsymbol{\Sigma}}_0 = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$ and $\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$.

Compute the estimated discriminant function $\hat{\delta}_k(\mathbf{x})$ for each k. Estimate the Bayes decision boundary between the two classes. Finally, construct a Bayes classifier to classify $\boldsymbol{x}_1^* = [2, 4]^T$ and $\boldsymbol{x}_2^* = [4, 1]^T$.

5. (15 pts) (*Discriminative Method*) Recall the K-Nearest Neighborhood (KNN) method which we covered in the lecture as a way to do regression. It can also be used as a way to construct a discriminative classifier. For simplicity, assume that we have one feature $x \in \mathbb{R}$ to predict an outcome variable $y = 0$ or 1.

   (a) (5 points) Given an integer K, write the algorithm or pseudo-code to describe how can you use the training data $\{x_i, y_i\}_{i=1}^n$ to estimate $P(y = 1|x = x_0)$, and hence construct a Bayes classifier $\hat{y}(x_0)$ for a new feature value x_0.

   (b) (5 points) Write the algorithm or pseudo-code to perform the leave-one-out CV to tune the hyper-parameter K based on the overall (testing) error rate (misclassification rate) for your classifier above.

(c) (5 points) The following R code aims to perform a 3-fold CV to estimate the error rate for the previous KNN classifier with $K = 5$. However, there is a lethal bug in the code that invalidates the estimated (testing) error rate. Explain what is the bug in the code, and in which way it bias the error rate.
Note: the function `knn_classification` was used to obtain the prediction for the testing set using the model fitted in the training set. You can assume this function is correct.

```
K_fold_CV_wrong <- function(K_KNN = 5, K_CV = 3, train_data){
  indx <- 1:nrow(train_data)
  # Split the indices into K folds
  kfolds <- createFolds(y = indx, k = K_CV, list = TRUE,
   returnTrain = FALSE)
  error_rate <- c()
  for (i in 1:K_CV) {
    # The training data in the i-th iteration
    selected_training_data <- train_data[-kfolds[[i]], , drop = F
   ]
    # The validation data in the i-th iteration
    selected_validation <- train_data
    # Compute the prediction
    i_th_pred <- knn_classification(k = K_KNN,
                          training = selected_training_data,
                          testing = selected_validation)
    # Compute the error rate in the i-th iteration
    new_error_rate <- mean(selected_validation$y == i_th_pred)
    error_rate <- c(new_error_rate, error_rate)
  }
  # Compute the mean error rate
  mean(error_rate)
}
```