Stat 437 HW4

Your Name (Your student ID)

General rule

Please show your work and submit your computer codes in order to get points. Providing correct answers without supporting details does not receive full credits. This HW covers

- Bayes classifier
- kNN classifier
- Discriminant analysis

For an assignment or project, you DO NOT have to submit your answers or reports using typesetting software. However, your answers must be well organized and well legible for grading. Please upload your answers in a document to the course space. Specifically, if you are not able to knit a .Rmd/.rmd file into an output file such as a .pdf, .doc, .docx or .html file that contains your codes, outputs from your codes, your interpretations on the outputs, and your answers in text (possibly with math expressions), please organize your codes, their outputs and your answers in a document in the format given below:

```
Problem or task or question ...
Codes ...
Outputs ...
Your interpretations ...
```

It is absolutely not OK to just submit your codes only. This will result in a considerable loss of points on your assignments or projects.

Conceptual exercises: I (Bayes classifier)

1. This exercise is on Bayes theorem and Bayes classifier.

1.1) State clearly the definition of the 0-1 loss function. Can this function be used in multi-class classification problems?

1.2) Let Y be the random variable for the class label of a random vector X, such that $Y \in \mathcal{G} = \{1, \ldots, K\}$ where $K \geq 2$ is the number of classes. Let $\hat{Y}$ be the estimated class label for X. Given the prior $\Pr(Y = k) = \pi_k, k \in \mathcal{G}$ on Class k and the conditional density $f_k(x)$ of X when it comes from Class k. Provide the formula to obtain the posterior $\Pr(Y = k|X = x)$, which is essentially the Bayes theorem. What is the Bayes classifier and how does it classify a new observation x_0 from X? Is the decision boundary of the Bayes classifier linear or quadratic in X? Explain (but do not have to mathematically prove) why the Bayes classifier minimizes the expected 0-1 loss. Note the a proof of the fact that the Bayes classifier minimizes the expected 0-1 loss is given in "LectureNotes4_notes.pdf". You should not copy and paste the proof. Instead, please provide the explanation based on your understanding of the proof.

1.3) If $K = 2$ in subquestion 1.2), what is the threshold value on $\Pr(Y = 1|X = x_0)$ that is used by the Bayes classifier to determine the class label for x_0? Suppose you use a different threshold value on $\Pr(Y = 1|X = x_0)$ to classify x_0, is the corresponding classifier still the Bayes classifier, and is the corresponding loss function still the 0-1 loss? Explain your answer. Provide a scenario where to classify an observation a different threshold value is more sensible than the threshold value used by the Bayes classifier.

1.4) If $K = 2$ in subquestion 1.2), $\pi_1 = 0.6$, $f_1(x) \sim \text{Gaussian}(0, 1)$ and $f_2(x) \sim \text{Gaussian}(2, 1)$ and $x_0 = 1.5$. Compute $\Pr(Y = 1|X = x_0)$ and use the Bayes classifier to classify x_0.

Conceptual exercises: II (k-NN classifier)

2. Given the training set $\mathcal{T}$ of n observations $(x_1, y_1), \ldots, (x_n, y_n)$, where y_i is the class label of observation x_i and $y_i \in \mathcal{G} = \{1, \ldots, K\}$ for $K \geq 2$, consider k-NN classifier, where k is the neighborhood size.

2.1) Describe how the decision boundary (such as its smoothness and shape) of k-NN classifier changes as k changes.

2.2) Explain why the training error of 1-NN classifier is 0. Provide an estimator of the test error of a classifier and explain why it can be used as such an estimator. Is it true that a large k leads to a k-NN classifier with smaller test error? Can the test error of a k-NN classifier be equal to the test error of the Bayes classifier? When k is large and k/n is small, what is a k-NN classifier approximately estimating?

2.3) When there are $K \geq 2$ classes, how does a k-NN classifier classify a test observation x_0?

2.4) When should data be standardized before applying a k-NN classifier? When standardizing data, do we standardize each observation or each feature?

2.5) Using your understanding of Example 3 in “LectureNotes4b_notes.pdf”, provide a step-by-step guide on how to choose an optimal k for k-NN classifier using cross-validation. You can provide such as guide in the form of “pseudocode” (see, e.g., https://en.wikipedia.org/wiki/Pseudocode for some details on pseudocode). Suppose the training set has few observations, can you still perform cross-validation in order to choose an optimal k? Explain your answer. (Hint: for the 2nd part, think about if having more observations helps better estimate test error.)

Conceptual exercises: III (Discriminant analysis)

3. Exercise 2 of Section 4.7 of the Text, which starts with “It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case.” (Helpful information on how to prove this is contained in the lecture video on LDA and “LectureNotes5b_notes.pdf”.)

4. Exercise 3 of Section 4.7 of the Text, which starts with “This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class specific mean vector and a class specific covariance matrix. We consider the simple case where p

$= 1$; i.e. there is only one feature." (Helpful information on how to prove this is contained in the lecture video on QDA and "LectureNotes5b_notes.pdf".)

5. Exercise 5 of Section 4.7 of the Text, which starts with "We now examine the differences between LDA and QDA." (Hint: for this question, you may also use information from Figures 4.9, 4.10 and 4.11 in the Text.)

6. Let Y be the random variable for the class label of a random vector $X \in \mathbb{R}^p$ (where p is the number of features), such that $Y \in \mathcal{G} = \{1, \ldots, K\}$ and $\Pr(Y = k) = \pi_k$ for Class k with $k \in \mathcal{G}$, where $K \geq 2$ is the number of classes. Consider the Gaussian mixture model such that the conditional density of X when it comes from Class k is $f_k(x) \sim \text{Gaussian}(\mu_k, \Sigma_k)$. Given the training set $\mathcal{T}$ of n observations $(x_1, y_1), \ldots, (x_n, y_n)$ on (X, Y), where y_i is the class label of observation x_i, do the following:

6.1) Provide the MLEs of π_k, μ_k and Σ_k for each $k \in \mathcal{G}$ respectively for the case where all Σ_k's are equal and for the case where not all Σ_k's are equal. When $p > n$, is the MLE of Σ_k still accurate? If not, recommend a different estimator for estimating Σ_k and provide details on this estimator.

6.2) Assume $p = 2$ and $K = 2$ and $k = 1$. For the density $f_k(x) \sim \text{Gaussian}(\mu_k, \Sigma_k)$, what shape do its contours take, and how does Σ_k control the shape of these contours? How do you check if the conditional density of X given that it comes from Class k is Gaussian?

6.3) Is it true that discriminant analysis will perform badly if the Gaussian assumption is violated? (Hint: for this question, you may also use the information provided by Figures 4.10 and 4.11 of the Text.) Let $X = (X_1, \ldots, X_p)^P$, i.e., X_1 up to X_p are the feature variables. Can discriminant analysis be applied to observations of X when some of $X_j, j = 1 \ldots, p$ is a discrete variable (such as a categorical variable)? Explain your answer.

6.4) What is a ROC curve, and what is AUC? How is AUC used to gauge the performance of a classifier? If you apply the same classifier, say, LDA or QDA under the same Gaussian mixture model, to two data sets that are independently generated from the same data generating process, i.e., that are independently generated from (X, Y) for classification problems, and obtain two ROC curves, would the two ROC curves be quite different? Explain your answer. When there are 3 or more classes, are the codes provided in the lecture notes able to obtain ROC curves and their AUC's for LDA and QDA?

6.5) Describe the key similarities and differences, respectively, between LDA and logistic regression. Provide a situation where discriminant analysis can still be sensibly applied but logistic regression is not well-defined.

Applied exercises: I (k-NN classifier)

7. Please refer to the NYC flight data `nycflights13` that has been discussed in the lecture notes and whose manual can be found at https://cran.r-project.org/web/packages/nycflights13/index.html. We will use `flights`, a tibble from `nycflights13`.

Please use `set.seed(123)` for the whole of this exercise. Randomly select from `flights` for each of the 3 `carrier` "UA", "AA" or "DL" 500 observations for the 3 features `dep_delay`, `arr_delay` and `distance`. Let us try to see if we can use the 3 features to identify if an observation belongs a specific carrier. The following tasks and questions are based on the extracted observations. Note that you need to remove rows with `na`'s from the extracted observations.

7.1) First, you need to standardize the features since they are on very different scales. Then randomly split the observations into a training set that contains 70% of the observations and a test set that contains the remaining observations.

7.2) Consider the observations as forming 3 classes that are determined by `carrier`. To the training set, apply 10 fold cross-validation to k-NN classifier with features `arr_delay`, `dep_delay`, and `distance` to determine the optimal k from the values $\{1, \ldots, 15\}$. Apply the optimal k-NN to the test set, provide the classification table and the overall error rate, and provide visualization of the classification results. Do you think the error rate is reasonable? Explain your answer. (Hint: you can follow the strategy provided by Example 3 in "LectureNotes4b_notes.pdf".)

7.3) Note that your have standardized the features `arr_delay`, `dep_delay`, and `distance`. However, with the unstandardized features, you would surely know that none of them follows a Gaussian distribution no matter with respect to which class (i.e., `carrier`) you look at their observations since these features have non-negative values (whereas a Gaussian distribution can generate negative values). Again, the 3 classes are determined by `carrier`. So, you will apply QDA based on the 3 standardized the features to the training set to train the model, and then apply the trained model to the test set (that contains the 3 standardized the features) to classify its observations.

(7.3a) First, check if the Gaussian assumption is satisfied. For this, note that if the standardized features `arr_delay`, `dep_delay`, and `distance` follow a trivariate Gaussian distribution for each individual class, then any pair among the 3 standardized features follows a bivariate Gaussian distribution for each individual class.

(7.3b) Apply the estimated (i.e., trained) QDA model to the test set, provide the estimated mixing proportion and estimated mean vector for each class, and provide the classification table. If you randomly pick an observation on the 3 standardized features, is it approximately equally likely to belong to each of the 3 carriers? (You do not need to mathematically prove your answer. However, think along this line: we are testing the equality of 3 population proportions, and each estimated population proportion is based on around 350 observations, which approximately can be done via a z-test since the central limit theorem is in effect.) How is the performance of QDA on this test set? Explain your answers.

(7.3c) Extract observations that are for "UA" or "DL" from the training set and the test set, respectively, to form a new training set and a new subset, so that there are now 2 classes "UA" and "DL". Apply QDA to the new training set and then apply the trained model to the new test set. Report the overall error rate on the test set, provide the ROC curve, and calculate the AUC. How is the performance of QDA on this test set? Explain your answer.

Applied exercises: II (Discriminant analysis)

8. The following is on software commands:

(8.1) What is the main cause of the message "Warning in `lda.default`(x, grouping, . . .): variables are collinear"? What is the main cause of the message "Error in `qda.default`(x, grouping, . . .) : some group is too small for 'qda' "?

(8.2) Provide details on the `list` that `predict{MASS}` returns.

(8.3) The arguments `gamma` and `lambda` of `rda{klaR}` are usually determined by cross-validation. Can they be set manually?

9. We will use the human cancer microarray data that were discussed in the lectures and are provided by the R library `ElemStatLearn` (available at https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/). Pick 3 cancer types "MELANOMA", "OVARIAN" and "RENAL", and randomly select the same set of 60 genes for each cancer type. Please use `set.seed(123)` for the whole of this exercise. Your analysis will be based on observations for these genes and cancer types.

9.1) Pick 2 features and visualize the observations using the 2 features. Do you think it is hard to classify the observations based on the amount of overlap among the 3 neighborhoods of observations in each of the 3 classes? Here "a neighborhood of observations in a class" is a "open disk that contains the observations in the class".

9.2) Apply LDA and report the classwise error rate for each cancer type.

9.3) Use the library `klaR`, and apply regularized discriminant analysis (RDA) by setting the arguments `gamma` and `lambda` of `rda{klaR}` manually so that the resulting classwise error rate for each cancer type is zero.

9.4) Obtain the estimated covariance matrices from the RDA and visualize them using the same strategy in Example 3 in "LectureNotes5c_notes.pdf". What can you say about the degree of dependence among these genes for each of the three cancer types? (Hint and caution: the class labels "MELANOMA", "OVARIAN" and "RENAL" will be ordered alphabetically by R. So, you need to keep track on which estimated covariance matrix is for which class. Otherwise, you will get wrong visualization.)