STATISTICS 446/646 - Exam two (Take home)
May 3rd, 2019

Name ___________________________________

Rules for take home exam: Please read the follwoing rules and confirm by signing the cover page that you have read and abide them while taking your exam.

(1) There are three pages including this cover page. You must turn in this signed cover page with exam; otherwise you will receive zero.

(2) You have exactly 24 hours to solve the exam. ABSOLUTELY NO EXTENSIONS.

(3) Exam must be submitted via eCampus using exam two submission link. The data and the exam with the submission link are posted inside the exam two submission folder at the left panel of eCampus.

(4) Please save your submission as a PDF, with format "lastname_firstname.pdf". Include all R code.

(5) The exam must be taken completely alone. Showing it or discussing it with anybody is forbidden, including (but not limited to) the other students in the course in current or previous years.

(6) You may use any publicly available material you want, including books, the internet, course materials at eCampus etc. (You are NOT allowed to submit questions to internet discussion groups, though!).

(7) Make an effort to make your submission clear and readable. Severe readability issues may be penalized by grade.

(8) Submit your code separately (or integrated into the solution or at the end of the solution) with comments and explanations. Even if the final result is wrong, the code may allow me to find the bug and award partial credit.

(9) There are 3 questions on the exam and total 50 points.

(10) GOOD LUCK!

On my honor, as an Aggie, I have neither given nor received unauthorized aid on this exam

Student's Signature_________________________________

Statistics 446/646: Statistical Bioinformatics

Exam two

May 3, Spring 2019

1. The "exam2_data1" data attached in the email are read counts from a next-gen sequencing (RNA-Seq) experiment on humans under two conditions (three healthy individuals and three diseased individuals).

   (a) Use the rlog function from the DESeq2 package to transform the read counts. Using the rlog-transformed counts.

      i. Carry out cluster analysis to explore structure among the assays (columns). Interpret the results. [**5 points**]

      ii. Carry out principal component analysis to explore structure among the assays (columns). Interpret the results. [**5 points**]

      iii. Use naïve Bayes to classify training individuals as diseased or healthy, using all 10,525 genes as predictor variables. Report a confusion matrix. What are the following estimated values: classification accuracy rate, sensitivity, specificity. You do not have to use cross validation or set aside any data for testing purposes. That said, how would you expect cross validation or test-set estimates of accuracy, sensitivity, and specificity to compare to the ones you reported? [**10 points**]

   (b) Use the DESeq function from the DESeq2 package to perform differential expression analysis. How many genes are selected at an FDR of 0.05 as being different between disease and healthy individuals? Use the plotCounts function to plot the normalized read counts (not raw read counts) for the top-ranked gene (the one with the smallest associated FDR estimate). [**15 points**]

2. Consider an RNA-seq experiment in which we are comparing a treatment group to a control group. Let Y_{ijk} be the read count for gene k in sample i of comparison group j, $i = 1, 2, ..., n_j$, $j = 1, ..., m$. What would be wrong with assuming that $Y_{ijk} \sim Poisson(\lambda_{jk})$? [**5 points**]

3. Context = "exam2_data2.csv." These are simulated, continuous -omics data. There are 2,000 features and 300 samples, with 100 samples each in three classes. In the data file, the first 75 samples are to be used as training samples for class A, the next 75 samples are to be used as training samples for class B, etc. So, the first 225 samples are training data. The last 75 samples are test data. The first 25 of these are for class A, the next 25 for class B, etc. [**10 points**]

```
[Hints: step 1: load the data
Y <- read.csv("exam2_data2.csv", row.names = 1)
m <- nrow(Y)
step 2:  Divide into training and testing data.
Y_train <- Y[, 1:225]
Y_test <- Y[, 226:300]
GRP_train <- factor(rep(c("A", "B", "C"), each = 75))
GRP_test <- factor(rep(c("A", "B", "C"), each = 25))
step 3: Arrange the training and test data into data frames.
Y_train_df <- data.frame("GRP" = GRP_train, t(Y_train))
Y_test_df <- data.frame("GRP" = GRP_test, t(Y_test))
Y_df <- rbind(Y_train_df, Y_test_df)]
```

(a) Classify with KNN. Consider values of $K = 1, 2, ..., 10$. Compute the test error for each value.

   i. Provide a plot with K on the x-axis and test accuracy on the y-axis. Interpret the plot.

   ii. Which value of K achieves the highest test accuracy, and what is that accuracy?