STAT 36-462/662, Final Exam, Due Sunday 11:59PM, May 13, 2018

This is an open book exam, and you can use textbooks and lecture notes. The full scores for the exam is 100 points. The exams contains two problems.

Problem 1 weights 15 points out of 100. Everybody should **finish the problem independently and submit it individually**, and if you have to discuss with your classmates, keep it to the minimum. Copy and paste the solution from one to the other is treated as cheating, and both parties will be penalized. Note that the instructor and the TA will only help you for clarifications, not to do the problem or provide hints for you. Please don't forget to put your Andrew ID, First Name, Last Name, and your signature.

Problem 2 weights 85 points out of 100. For this problem, you are supposed to form a team of at most 4 students, and **each team is only required to submit one copy of the solution**. In the solution, please include the size of the team, and also for each team member, please include the Andrew ID, First Name, Last Name, and the signature, arranged in the alphabetical order. The two problems should be **submitted separately to the head TA**.

Problem 1. (15 pts = 5 + 5 + 5). Given training samples (x_i, y_i), $i = 1, 2, \ldots, n$, where $y_i \in \{1, 2\}$ and $x_i \in R^p$ $(p > 2)$, and a fresh sample $x \in R^p$, we suppose $\pi_1 = \pi_2 = 1/2$, and that

$$f(x|y = 1) \sim N(\mu_1, \Sigma), \qquad \text{and} \qquad f(x|y = 2) \sim N(\mu_2, \Sigma).$$

We are interested in predicting the class label associated with x.

- (a). Show that Fisher's LDA rule reduces to

$$y \begin{cases} = 2, & \text{if } \hat{\delta}(x) > 0, \\ = 1, & \text{if } \hat{\delta}(x) \le 0, \end{cases}$$

  where

$$\hat{\delta}(x) = (\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} x - \frac{1}{2}(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2),$$

  where $(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma})$ are as in class.

- (b). We now assume (μ_1, μ_2, Σ) as known. Let $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$. The $p \times p$ between-class variance matrix B introduced in Lecture 22 now reduces to

$$B = \frac{1}{2}[(\mu_1 - \bar{\mu})(\mu_1 - \bar{\mu})^T + (\mu_2 - \bar{\mu})(\mu_2 - \bar{\mu})^T].$$

  We are interested in finding a nonzero vector $a \in R^p$ to maximize

$$\frac{a^T B a}{a^T \Sigma a}. \tag{1}$$

  Denote the solution by a_0. Show that $\Sigma^{1/2} a_0$ is an (unscaled) eigenvector of $\Sigma^{-1/2} B \Sigma^{-1/2}$, a_0 is an (unscaled) eigenvector of $\Sigma^{-1} B$, and $a_0 \propto \Sigma^{-1}(\mu_2 - \mu_1)$.

- (c). We continue to assume (μ_1, μ_2, Σ) as known. We now project the fresh data vector x from p-dimension to 1 dimension by (where a_0 is as in (b))

$$x \qquad \mapsto \qquad a_0^T x.$$

  Write down the distribution of $(a_0^T x|y = 1)$ and $(a_0^T x|y = 2)$. Derive the Fisher's LDA classification rule applied to $a_0^T x$. Compare this rule with that in (a) (for the rule in (a), you can replace $(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma})$ by (μ_1, μ_2, Σ) for the latter are known). Justify your answers.

Problem 2. (85 pts). The data for your final project is for 133 patients from two classes (a control class and a case class), measured on the same set of 250 genes (these are the genes selected from a much larger set of genes). The data set can be downloaded on canvas from the newly created folder `FinalExam/TrainData.RData`, where `trainX` is a data frame with 250 genes and 133 patients, and trainY is a 0-1 label vector where 0 represents control and 1 represents case. **Note that I have kept to myself a test data set for about** 50 **patients**. The test data set is not accessible to you, but the TA will use it to evaluate your algorithms. We are interested in both unsupervised learning and supervised learning.

1. **Unsupervised learning**. In this problem, you are not allowed to use the class labels provided to you when you learn the data (you can however use the class labels to evaluate your clustering errors after you have built your clustering methods; see below).

   (a) (10) Is there any interesting structure present in the data? Since the dimension is high, you may apply some dimension reduction methods we learned in the class. This will help you in visualization and interpretation.

   This question is intentionally vague. If you do't find anything interesting, then describe what you tried, and show that there isn't much visible structure. Data mining isn't about continuing to manipulate the data until you get an answer.

   (b) (30) Which clustering methods seem to work better on your data sets? Describe what you have tried, and present your results, especially the Hamming clustering errors. In particular, does the dimension reduction step suggested in (a) help you in clustering?

   Note that in addition to error rates, we also care interpretability. Pick the clustering method you think that is the "best" for your data set, submit the R code to the TA, and justify why you think the method is the "best". The TA will apply your R code to the **whole data set** (note that I have held out the data for another 48 samples) and compute the Hamming clustering errors. You will be taken 10 points off if the TA finds your code not working (provided that it is your coding error).

2. **Supervised learning**. (45 pts). In this problem, you can use the class labels provided to you. How did you make your predictions? Describe this process in detail. Again, you can use any of the classification techniques that we learned in the course, or any other techniques as long as they are adequately described. What assumptions, if any, are being made by using these techniques? If there were tuning parameters, how did you pick their values?

   Pick the "best" classifier and submit the R code to the TA, justify why you think the classifier is the best. The TA will apply your classifier to the **test data set** (which is not accessible to you) and compute the Hamming classification error. You will be taken 10 points off if your R code does not work (provided that it is your fault).

TA mentors. Each group will be assigned a TA mentor. Ask them for help! They have valuable expertise and experience.

R code. For each of the two problems, you are required to submit a piece of R code. Make sure the R code runs well, and add necessary comments in the R code (without affecting its running).

Write-up. Additional to the R code, you are also required to submit a write-up. The write-up should be a polished report, with figures and snippets of R code as you deem helpful. You don't need to submit your (this part of) R code in its entirety. Your report should have the following sections (you can of course add subsections if you want): Introduction (description and the motivation of the problems, organization of your report, etc. Please keep it short), unsupervised learning, supervised learning. The report should be no more than 7 pages.

Grading. Your points depend on three components: your write-up (clarify, completeness, etc.), the grasp of the course material you demonstrate (e.g., interpretation of your results, rationale for picking certain methods to try, correctness of your approach and code), and your clustering/classification error rates.

For example, for unsupervised learning, a prefect score is 40 points. We will take off points if (a) your presentation is poor, (b) your clustering/classification errors are substantially poorer than expected, and (c) the approaches you have taken have significant errors or not adequately justified or interpreted, (d) the set of methods you have investigated are far fewer than expected.

Bonus points. For unsupervised learning, the three teams that have the smallest clustering error rates on the whole data set will receive a 3 bonus points. For supervised learning, the three teams that have the smallest classification error rates on the test data set will also receive a 3 bonus points. In either case, if there is a tie, the tie breaker will be your presentation (e.g., clarity of your presentation and justifications for why pick the method you recommended). This means in theory, a team can receive 106 points out of 100!

Cheating. Don't cheat. We know that there are ways to cheat on this final project. If we suspect you of cheating (e.g., if you have a remarkably low misclassification rate, but your method is not really statistically motivated), then we reserve the right to give you a 0.