Your name: _____

Your SUNet ID: _____

Exam rules:

- You have until 4:00 PM July 22, 2021 to complete the exam and submit it to Gradescope.
- Following the updated BJA guidelines for open book exams, you are allowed to use your textbook, course slides, notes, and the internet in completing this exam.
- A Cheat Sheet is provided at the end of the exam.
- Please show your work and justify your answers.

Problem	Points	Max
1		20
2		10
3		10
4		10
5		20
6		10
7		20
Total		100

1. We define a new kind of discriminant analysis for a classification problem with a binary response. The classes have prior probabilities π_0 and π_1. Given the class, k, the conditional probability of the inputs $X_1, \ldots, X_p$ is multivariate normal with a class-dependent mean μ_k and covariance matrix $\sigma_k \mathbf{\Sigma}$. The matrix $\mathbf{\Sigma}$ is common to both classes and σ_k is a class-dependent constant. All parameters, π_k, μ_k, σ_k, for each class, as well as Σ, are set to their Maximum Likelihood estimates.

   (a) **[10 points]** Provide an equation describing the classifier's decision boundary or discriminant. What would the boundary look like?

   Note that this is a special case of Quadratic Discriminant Analysis. The covariance matrix for the inputs in each class may be different, but they are constrained to be related by scaling by a constant. The discriminant is the same as in QDA, but with this additional constraint. It is described by equating the objective functions for the two classes $\delta_1(x) = \delta_0(x)$, in this case

$$\log \pi_1 - \frac{1}{2}\sigma_1^{-1}\mu_1^T\mathbf{\Sigma}^{-1}\mu_1 + \sigma_1^{-1}x^T\mathbf{\Sigma}^{-1}\mu_1 - \frac{1}{2}\sigma_1^{-1}x^T\mathbf{\Sigma}^{-1}x - \frac{1}{2}\log|\sigma_1\mathbf{\Sigma}| =$$

$$\log \pi_0 - \frac{1}{2}\sigma_0^{-1}\mu_0^T\mathbf{\Sigma}^{-1}\mu_0 + \sigma_0^{-1}x^T\mathbf{\Sigma}^{-1}\mu_0 - \frac{1}{2}\sigma_0^{-1}x^T\mathbf{\Sigma}^{-1}x - \frac{1}{2}\log|\sigma_0\mathbf{\Sigma}|.$$

   This is still a quadratic equation in x.

   (b) **[5 points]** Why might this classifier be preferable to Linear Discriminant Analysis?

   The quadratic boundaries are more flexible than linear boundaries. This model allows you to fit cases in which different classes have different spreads.

   (c) **[5 points]** Why might this classifier be preferable to Quadratic Discriminant Analysis?

   By constraining the relationship between the covariance matrices for different classes, we have to estimate much fewer parameters from the data. This reduces the variance of the classification, which could lower the test error.

2. **[10 points]** Compare leave-one-out cross validation to 10-fold cross validation, with reference to the bias-variance tradeoff.

   The training sets in 10-fold cross validation are smaller than the full data, so the test error estimates tend to be biased upward. Leave-one-out cross validation estimates are nearly unbiased. On the other hand, leave-one-out cross validation yields estimates with greater variance, since the training sample hardly changes.

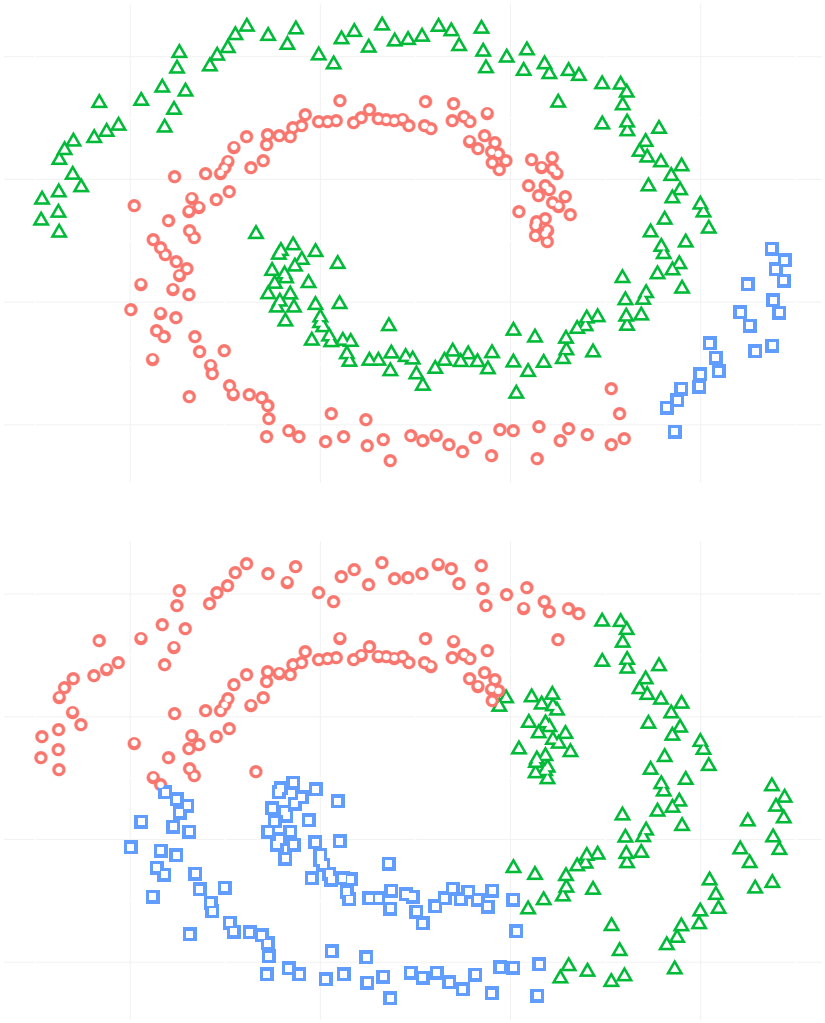
3. **[10 points]** A total of n samples were simulated from the following distribution

$$X_1, X_2, X_3, X_4 \sim \mathcal{N}(0, 1) \text{ i.i.d.}$$
$$Y = X_1 + 2X_2 + X_3^3 + X_1X_4 + \epsilon,$$

   where f is non-linear. Consider the following regression methods for Y: linear regression with predictors X_1, X_2, X_3, and X_4, and 3-nearest neighbors regression. On the same plot, sketch a plausible learning curve for each method. A learning curve for regression shows the average test MSE as a function of n. Explain your reasoning.

   When n is small, it is likely that the linear model would dominate 3-nearest neighbors, as this model suffers from the curse of dimensionality. However, with n large enough, 3-nearest neighbors would achieve a lower test error, because it is a non-parametric model capable of capturing any regression function. Linear regression will achieve a test error that is strictly larger than the irreducible error.

4. **[10 points]** The clusterings below were produced by single-linkage hierarchical clustering and k-means clustering. Determine which one is which and explain your reasoning.



The first clustering is generated by single-linkage hierarchical clustering, because it is easy to see that some triangles are nearest to the center of all circles.

5. We apply the Bootstrap to a dataset with n distinct observations $x_1, \ldots, x_n$.

   (a) **[10 points]** What is the probability that the j^{th} observation is included in a specific bootstrap sample?

   Let the Bootstrap sample of n observations be P_n^b. The probability that $x_j \notin P_n^b$ for a single draw is $(n-1)/n$. Since every observation in the Bootstrap sample is sampled independently,

   $$P(x_j \notin P_n^b) = \prod_{i=1}^{n} P(x_j \notin P_n^b) = \left(\frac{n-1}{n}\right)^n.$$

   The probability that the jth observation is in the Bootstrap sample is

   $$1 - P(x_j \notin P_n^b) = 1 - \left(\frac{n-1}{n}\right)^n.$$

   (b) **[10 points]** What is the expected value for the fraction of distinct observations in a specific bootstrap sample (i.e. the number of distinct observations from $x_1, \ldots, x_n$ divided by n). Does this expectation converge as n grows large? Hints: (i) use the probability from part (a), (ii) $\lim_{n\to\infty}(1-1/n)^n = e^{-1}$.

   Letting $\mathbb{I}(x_j \in P_n^b)$ be an indicator variable for the event that the jth observation is in the Bootstrap sample, the number of distinct observations in the sample is $\sum_{j=1}^{n} \mathbb{I}(x_j \in P_n^b)$. By linearity of expectation:

   $$\begin{aligned} E\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(x_j \in P_n^b)\right] &= \frac{1}{n}\sum_{j=1}^{n} E[\mathbb{I}(x_j \in P_n^b)] \\ &= \frac{1}{n}\sum_{j=1}^{n} P(x_j \in P_n^b) \\ &= \frac{1}{n}\sum_{j=1}^{n} \left\{1 - \left(\frac{n-1}{n}\right)^n\right\} \\ &= 1 - \left(\frac{n-1}{n}\right)^n = 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \quad \text{as } n \to \infty. \end{aligned}$$

6. **[10 points]** A group of 33 people were asked to report their happiness on a scale from 0 to 20. We apply a linear model with an intercept to regress happiness onto 2 predictors, the yearly income and the amount of money paid in taxes last year.

   The t-statistics for income and taxes have corresponding p-values of 0.14 and 0.52, respectively. The RSS of the model is 30 and the sample variance of the happiness is 1.5.

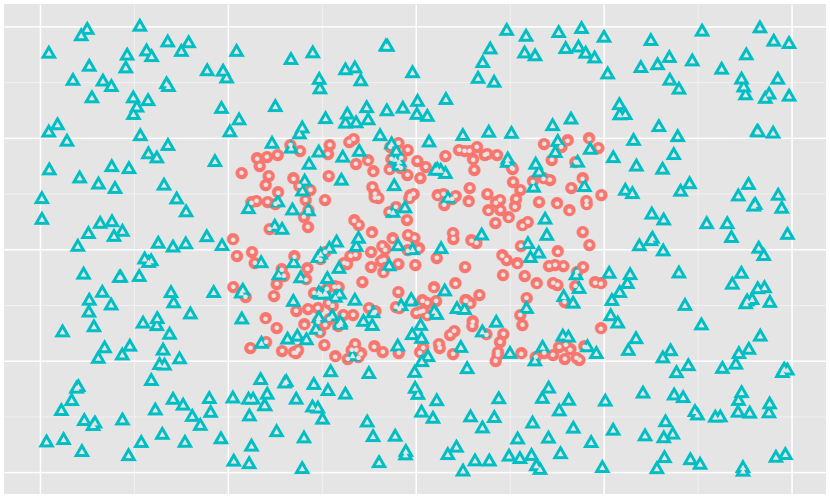
   What would you conclude about the relationship between happiness, income, and tax contributions?

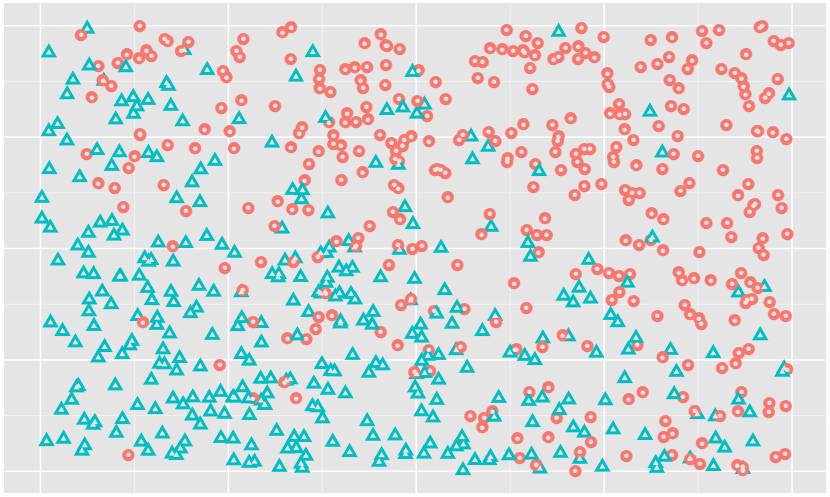
   We perform an F-test for the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. We have $RSS_0 = 1.5(n-1) = 1.5 \cdot 32 = 48$. Then, the F-statistic is:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} = \frac{(48-30)\cdot 30}{30 \cdot 2} = 9.$$

   The F-test would reject this hypothesis at the 1% significance level. This would suggest that the two predictors are important, even though their individual t-tests would suggest otherwise. This is probably due to collinearity, since the amount paid in taxes every year correlates with the income.

7. (a) **[10 points]** Identify which classifier among k-nearest neighbors with $k = 15$ and logistic regression would be more appropriate for each dataset below. Explain how one might adjust the True Positive rate of each method.





Note: Red circles are negative and blue triangles are positive.

Top: Since the decision boundary seems very non-linear and there are only 2 predictors, I would use a k-nearest neighbors algorithm. The k-nearest neighbors algorithm classifies to the positive class if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{1}{n} \sum_{i \in N_k(x)} \mathbf{1}(y_i = +)$$

is greater than 0.5. To increase the rate of True Positives, we could lower this threshold.

n.b. I will also accept (as correct) a linear model with indicator functions as presented in lecture 8.

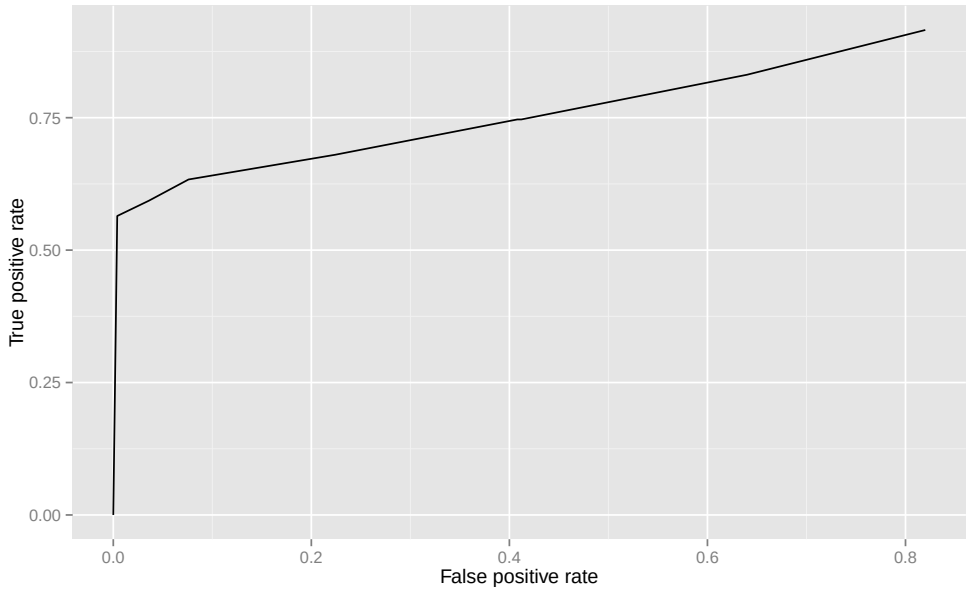
n.b. Regarding changing k to increase the True Positive rate, it is not clear that changing k will do this as it'll depend upon how many observations there are in each class. Assume for now that we have the same number of positive and negative cases. Notice that (i) the observations are distributed approximately uniformly, and (ii) the blue triangles are more spread out than the red circles. Now, for the blue triangles outside the inner rectangle (created by the circles), increasing k can only reduce the TPR since you can only search inward (towards the circles) to get the sufficient neighbors that you'd need which would include more negatives in your votes and cause you to vote negative. If you're inside the smaller rectangle, you still can't easily get more positives since the circles are more densely packed (resulting in more negatives picked anyway). Consequently, you actually want to pick smaller k to get a higher True Positive rate.

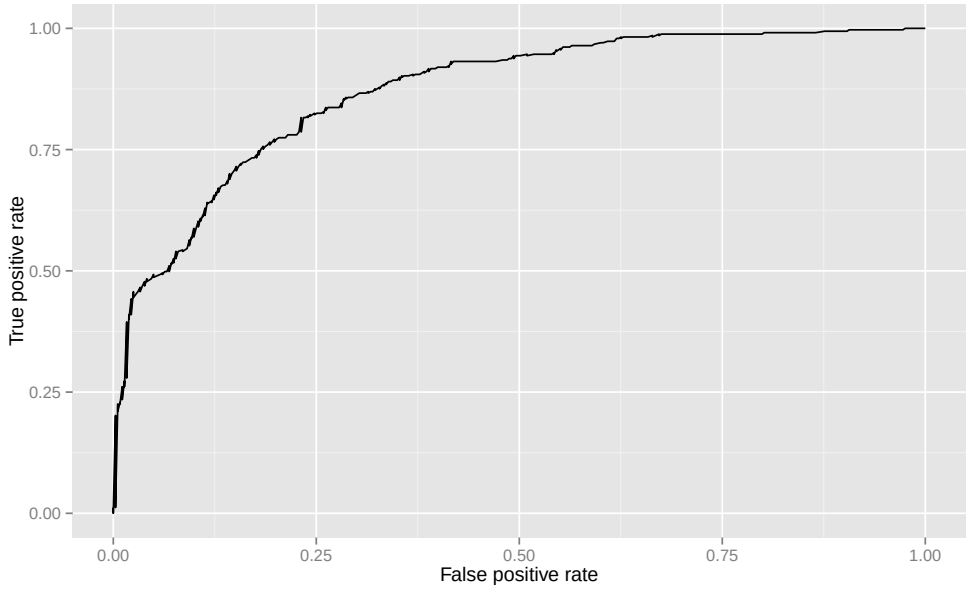
Bottom: The decision boundary seems linear or close to linear, so I would use LDA or logistic regression. Logistic regression assigns to positive if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{e^{X \cdot \hat{\beta}}}{1 + e^{X \cdot \hat{\beta}}}$$

is greater than 0.5. To increase the rate of True Positives, we could lower this threshold.

(b) **[10 points]** Each of the ROC curves below corresponds to one of the datasets in part (a). In each case, we applied the optimal classifier among k-nearest neighbors and logistic regression. Match each ROC curve to its corresponding dataset and explain your reasoning.





The top ROC curve corresponds to the first dataset. If the threshold is very small, we classify everything as positive (blue triangle), which is the top right corner of the plot. As we increase the threshold, we start to classify some red points inside the square as red, whose neighbors are mostly red, and some blue points inside the square as red as well. This would decrease the true positive rate and the false negative rate a bit. The elbow corresponds to the point in which all points inside the square are classified as red, at which point the True positive rate is still above 0.5. Then, sharply, all points are classified as red, bringing the false positive and true positive rates to zero.

The bottom ROC curve corresponds to the second dataset, where as we increase the threshold, the true positive and false positive rates decrease gradually — the decision boundary moves from the red region to the blue region.

Cheat sheet

The sample variance of $x_1, \ldots, x_n$ is:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

The residual sum of squares for a regression model is:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

t-test:

The t-statistic for hypothesis $H_0 : \beta_i = 0$ is

$$t = \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}$$

F-test:

The F-statistic for hypothesis $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ is

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)},$$

where RSS_0 is the residual sum of squares for the null model H_0, and RSS is the residual sum of squares for the full model with all predictors. Asymptotically, the F-statistic has the F-distribution with degrees of freedom $d_1 = q$ and $d_2 = n - p - 1$.

Minimum F-statistic to reject H_0 at a significance level $\alpha = 0.01$

		d_1			
		1	2	3	4
d_2	1	4052.181	4999.500	5403.352	5624.583
	10	10.044	7.559	6.552	5.994
	20	8.096	5.849	4.938	4.431
	30	7.562	5.390	4.510	4.018
	120	6.851	4.787	3.949	3.480

Logistic regression:

Logistic regression assigns to positive if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{e^{X\cdot\hat{\beta}}}{1 + e^{X\cdot\hat{\beta}}}$$

LDA:

The log-posterior of class k given an input x is:

$$C + \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + x^T \mathbf{\Sigma}^{-1}\mu_k$$

where C is a constant which does not depend on k.

QDA:

The log-posterior of class k given an input x in QDA is:

$$C + \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}_k^{-1}\mu_k + x^T \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}x^T \mathbf{\Sigma}_k^{-1}x - \frac{1}{2}\log|\mathbf{\Sigma}_k|$$

where C is a constant which does not depend on k.