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?

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

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

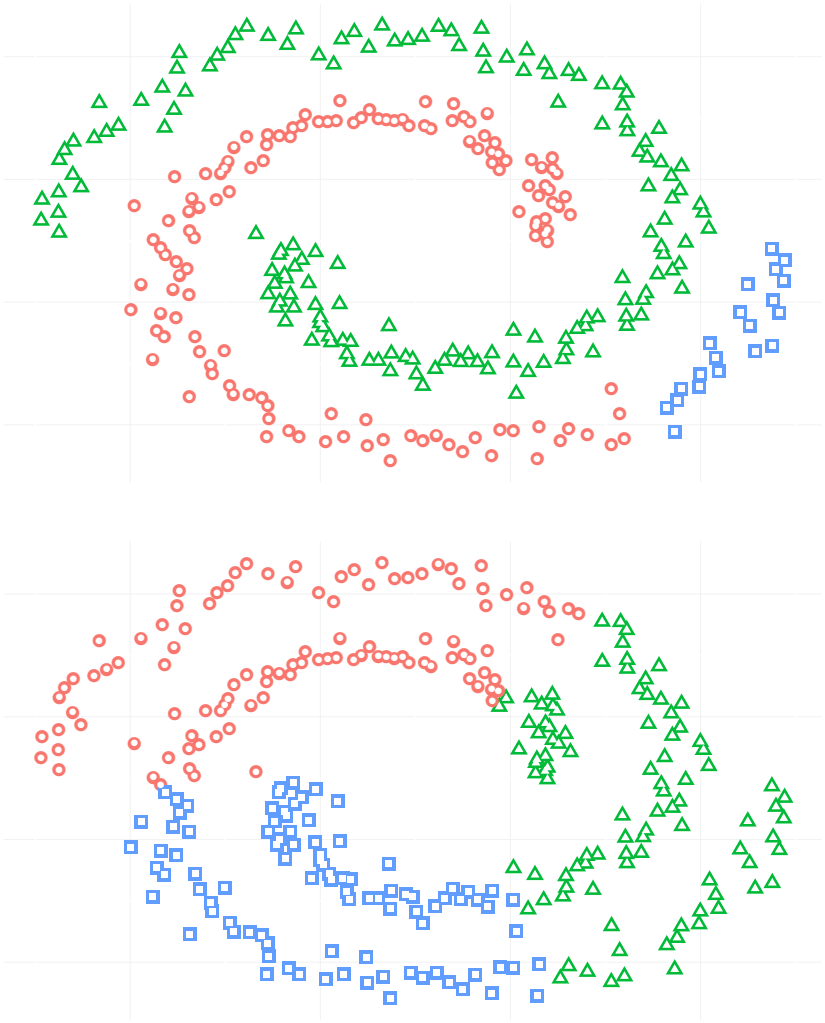
2. **[10 points]** Compare leave-one-out cross validation to 10-fold cross validation, with reference to the bias-variance tradeoff.

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.

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.



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?

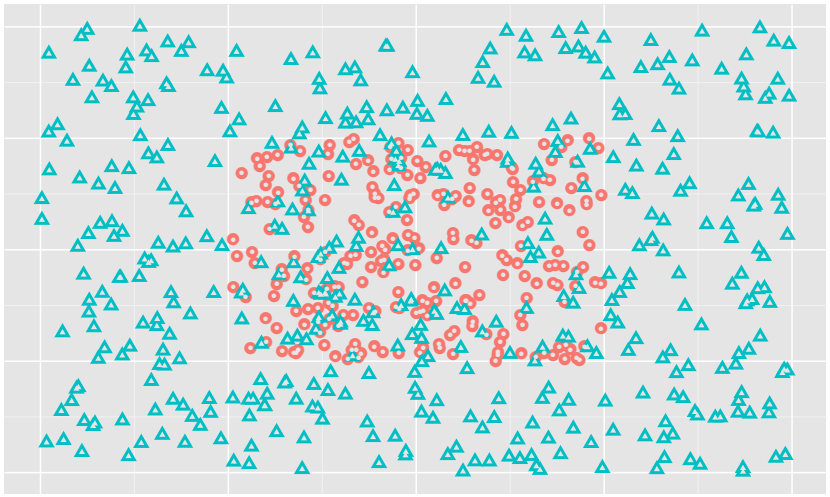
   (b) **[10 points]** What is the expected value for the fraction of distinct observations in a specific bootstrap sample (i.e. the number of distinct observations from $x_1, \ldots, x_n$ divided by n). Does this expectation converge as n grows large? Hints: (i) use the probability from part (a), (ii)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

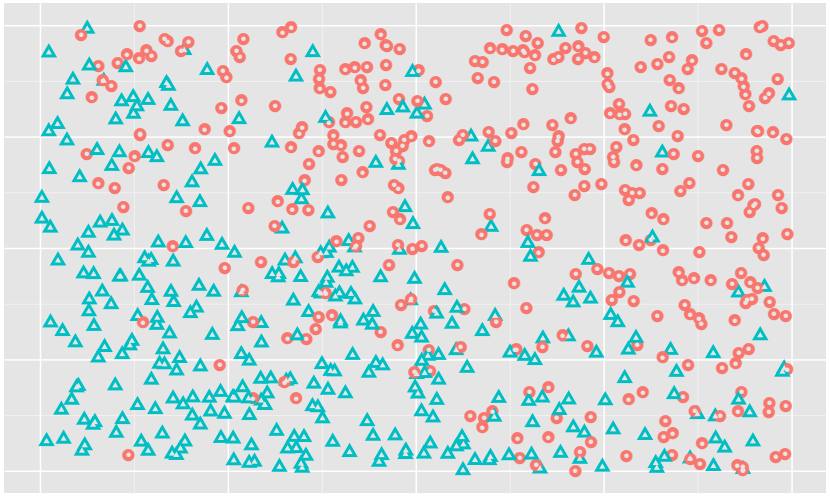
6. **[10 points]** A group of 33 people were asked to report their happiness on a scale from 0 to 20. We apply a linear model with an intercept to regress happiness onto 2 predictors, the yearly income and the amount of money paid in taxes last year.

   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?

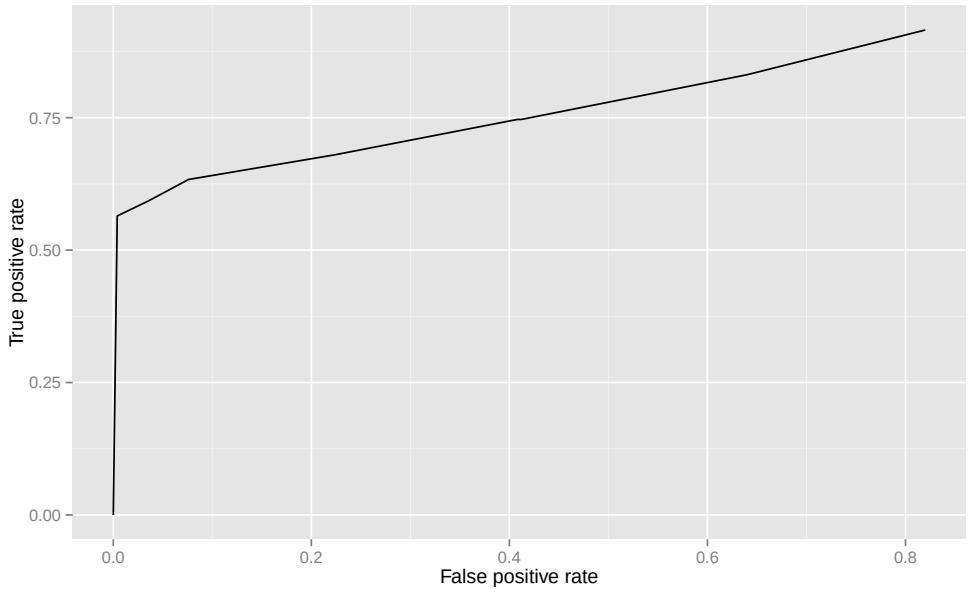
7. (a) **[10 points]** Identify which classifier among k-nearest neighbors with $k = 15$ and logistic regression would be more appropriate for each dataset below. Explain how one might adjust the True Positive rate of each method.

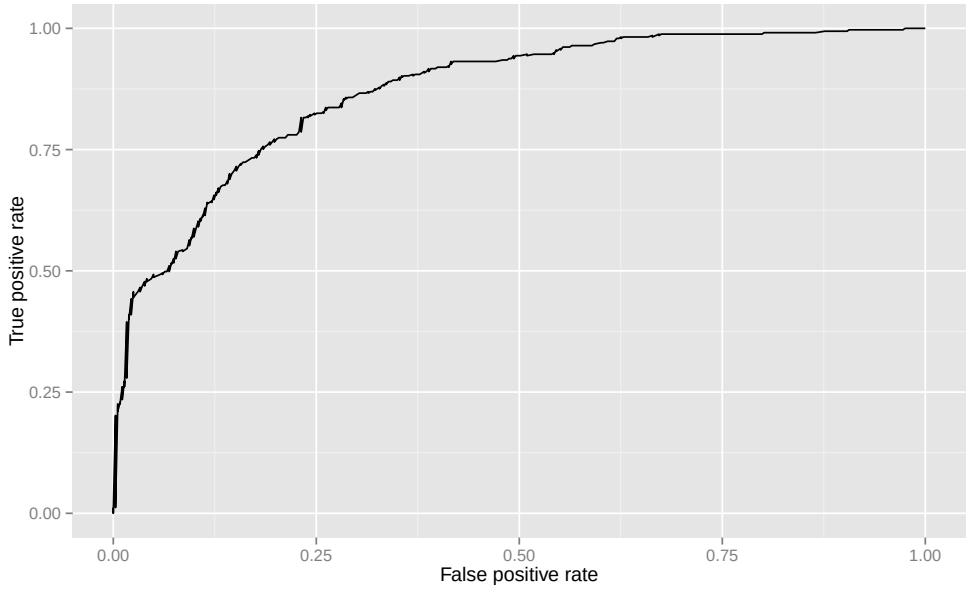




Note: Red circles are negative and blue triangles are positive.

(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.





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.