Your name: ____

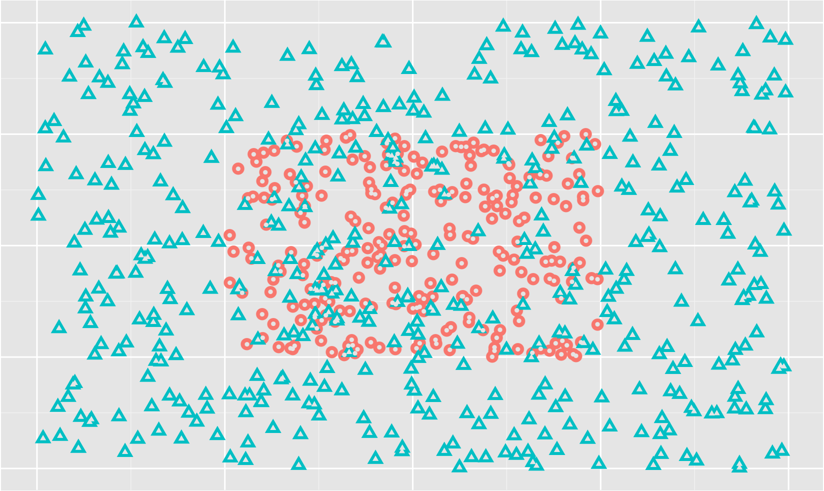
Your SUNet ID: ____

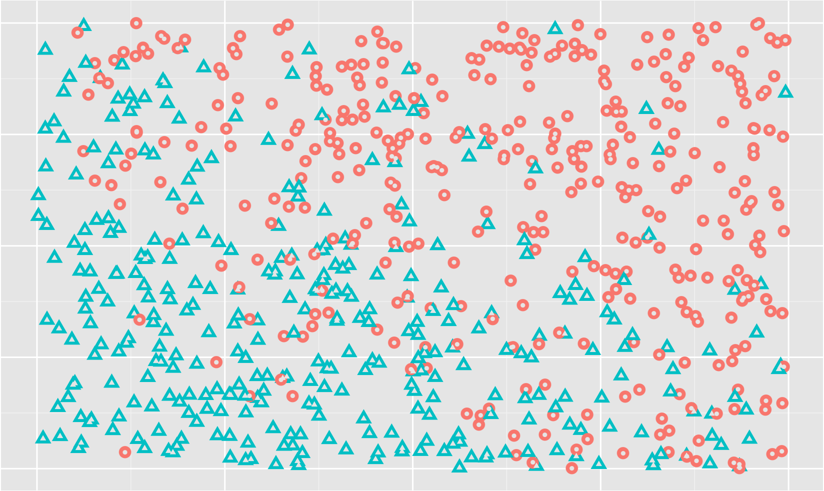
Exam rules:

- You have 50 minutes to complete the exam.
- You are not allowed to consult books or notes, or to use calculator or cell phone. If you must use a computer to type your solutions, you are not allowed to use any software aside from a Word processor or LaTeX.
- Please show your work and justify your answers.
- **SCPD students:** If you are taking the exam remotely, please return your solutions along with a routing form, signed by your proctor, by 2 pm PST on Tuesday, October 28. You can email a PDF or Word file to scpd-distribution@lists.stanford.edu or fax the solutions to 650-736-1266.

Problem	Points
1	
2	
3	
4	
Total	

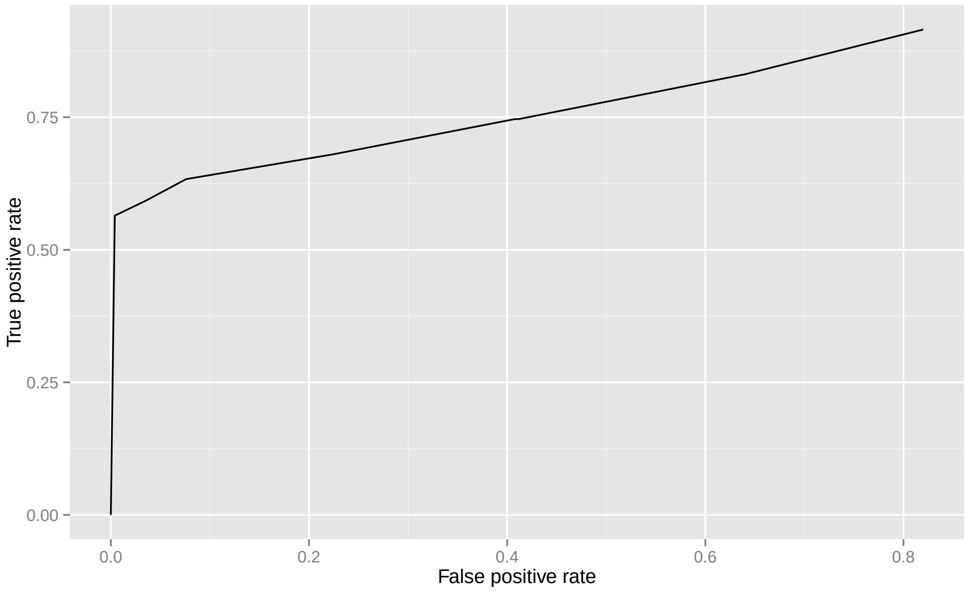
1. (a) Identify which classifier among k-nearest neighbors with $k = 15$ and logistic regression would be more appropriate for each dataset below. Explain how one might adjust the True Positive rate of each method.

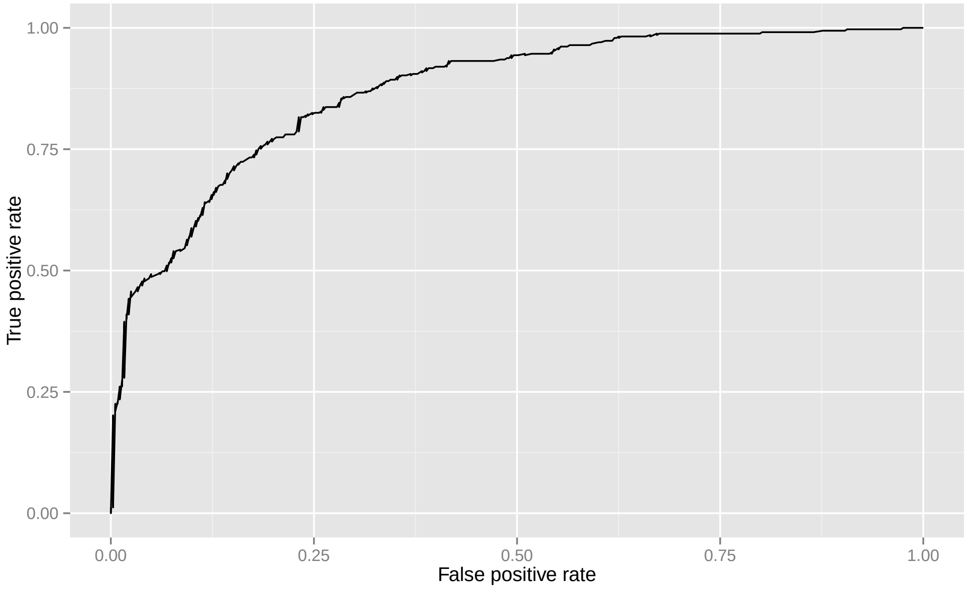




Note: Red circles are negative and blue triangles are positive.

(b) 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.




1.00
0.75
0.50
0.25
0.00
True positive rate
0.00
0.25
0.50
0.75
1.00
False positive rate

2. Two distances, d and d', are related by a monotone transformation:

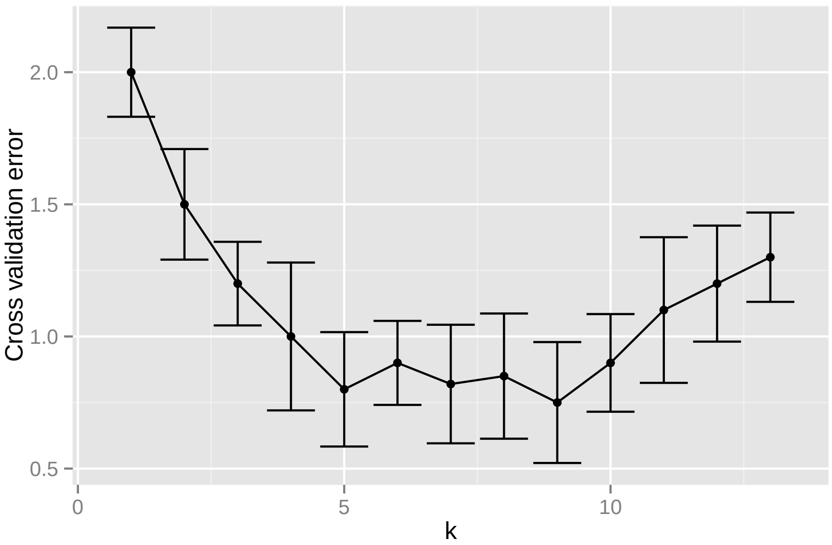
$$d'(a,b) = f(d(a,b))$$

which satisfies $f(x) \geq f(y)$ if $x \geq y$.

(a) Prove that the single linkage hierarchical clustering with k clusters is the same under d and d'.

(b) Prove that the complete linkage hierarchical clustering with k clusters is the same under d and d'.

3. State and explain the one standard error rule for model selection using 10-fold cross validation. Apply it to select the optimal number of nearest neighbors in the plot below, which shows the cross-validation error and one standard error intervals as a function of k.



4. 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 curves for each method. A learning curve for regression shows the average test MSE as a function of n. Explain your reasoning.