Your name: ____

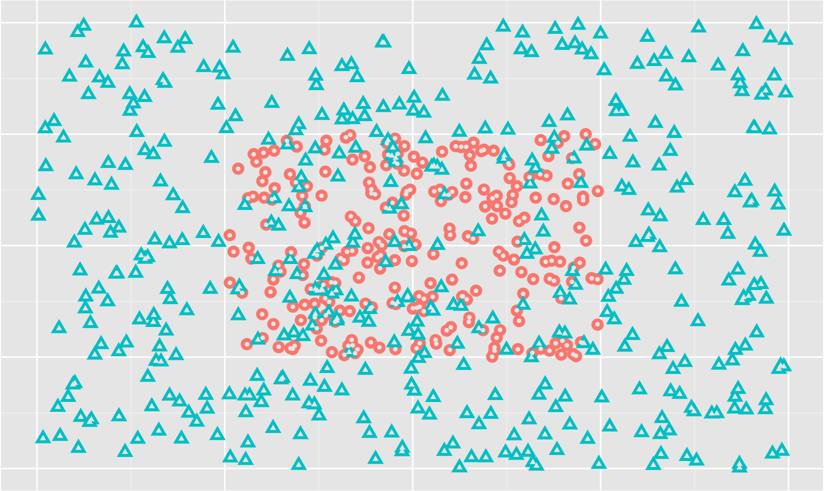
Your SUNet ID: ____

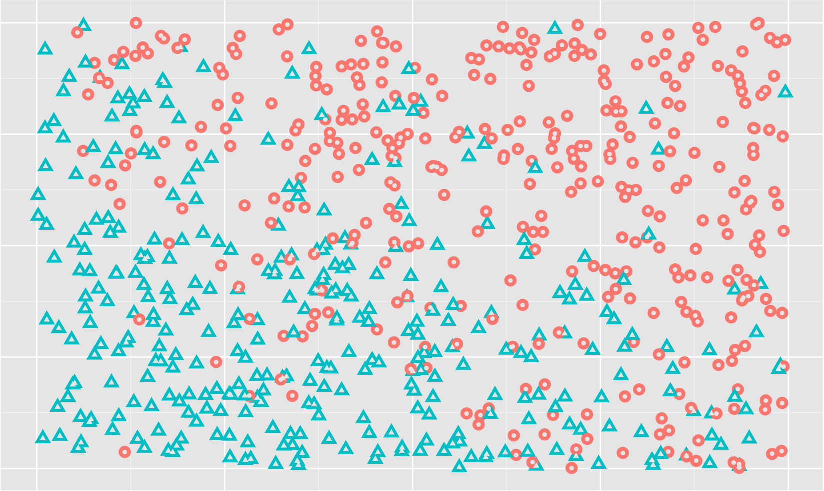
Exam rules:

- 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.

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

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

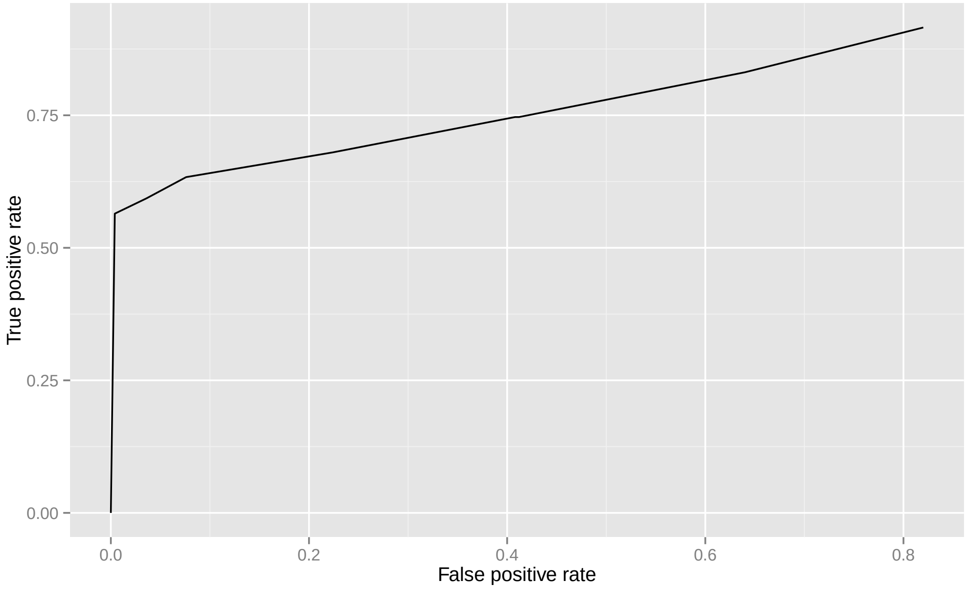
is greater than 0.5. To increase the rate of True Positives, we could lower this threshold.

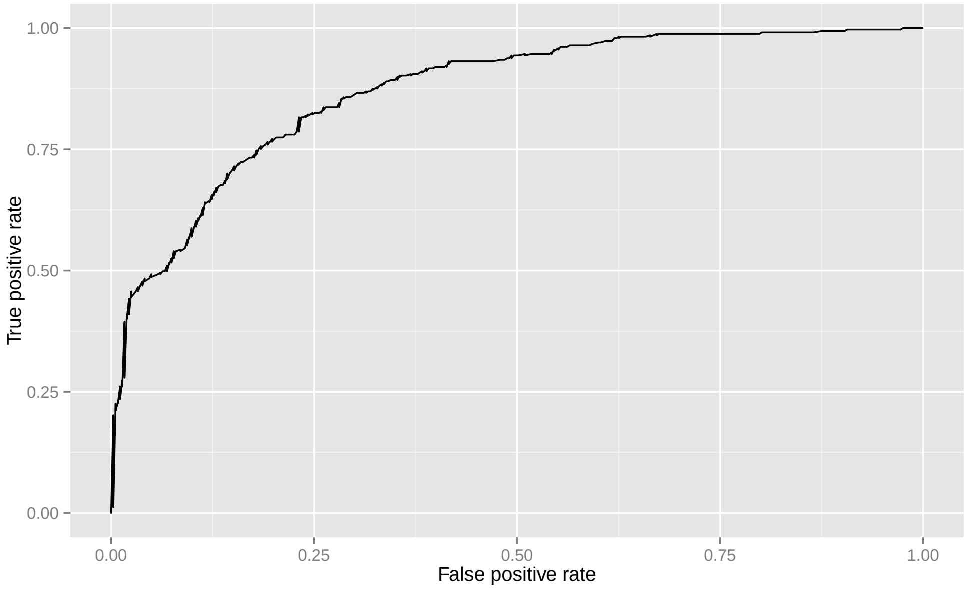
Bottom: The decision boundary seems linear or close to linear, so I would use LDA or logistic regression. Logistic regression assigns to positive if the estimated conditional probability

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

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

(b) 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.

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'.

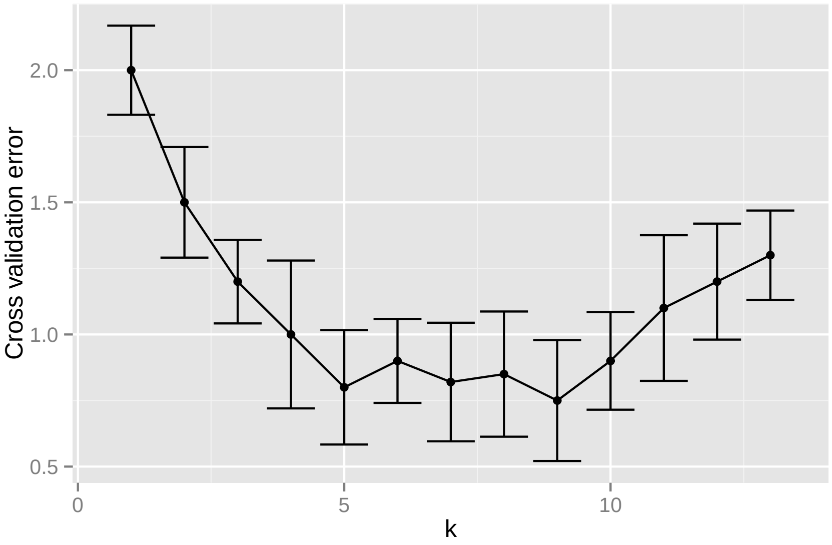
At each step of an agglomerative clustering algorithm, we join the two clusters that are closest together. Suppose at some level in the dendrogram, the clusters are the same under d and d'. Let A and B be two clusters, and (a,b) be the pair of samples that are closest together under d, with $a \in A$ and $b \in B$. Since d' is a monotone transformation of d, the pair of points in A and B that are closest together under d' will also be (a,b). The single-linkage distance between clusters A and B is then $d(a,b)$ in the first case, and $d'(a,b)$ in the second case.

Now, suppose that A^* and B^* are the two clusters that are closest together under d. By monotonocity again, A^* and B^* will be the most proximal clusters under d'. This implies that the next pair of clusters to be joined in the dendrogram is the same under both distances. By induction, the two dendrograms have the same structure, and the clustering with k clusters will be identical.

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

The proof follows the same argument as above. The complete-linkage distance between clusters A and B is just the distance between two samples a and b, and by monotonicity, these will be the same two samples under d and d'. Then, at every step of the agglomerative algorithm we join the two closest clusters, and because of the previous fact and monotonicity, this pair of clusters is always the same under d' and d. Hence, the dendrograms have the same structure and the clusterings with k clusters are identical.

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.



The one-standard error rule states we should choose the simplest model whose error lies within a standard error of the minimum error. The minimum error in the plot above is achieved at $k = 9$. The flexibility or variance of k-nearest neighbors decreases with k, so we would have to choose a model with $k \geq 9$. The model with $k = 10$ is the only model whose error lies within a standard error of the minimum error, so we would pick $k = 10$.

4. A total of n samples were simulated from the following distribution

$$X_1, X_2, X_3, X_4 \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

$$Y = X_1 + 2X_2 + X_3^3 + X_1 X_4 + \epsilon,$$

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.

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.