Your name: _____

Your SUNet ID: _____

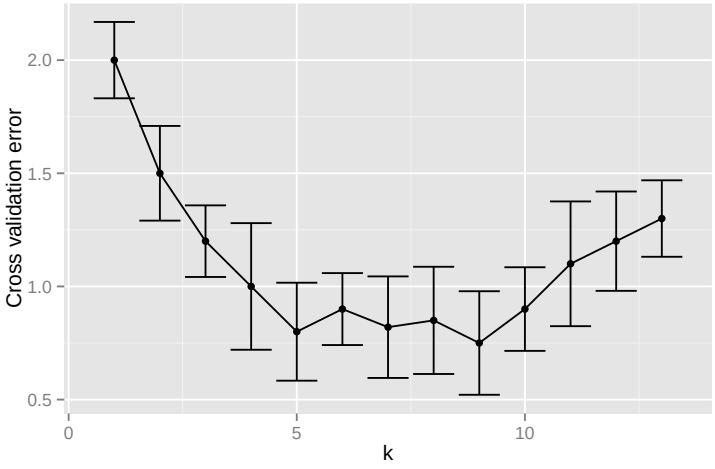
Exam rules:

- You have until 4:00 PM July 22, 2021 to complete the exam and submit it to Gradescope.
- You are only allowed to consult your course textbooks. You are not allowed to consult other material, textbooks, computers, cell phones, the internet, or other people. 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.
- A Cheat Sheet is provided at the end of the exam.
- Please show your work and justify your answers.

Problem	Points
1	
2	
3	
4	
5	
6	
7	
Total	

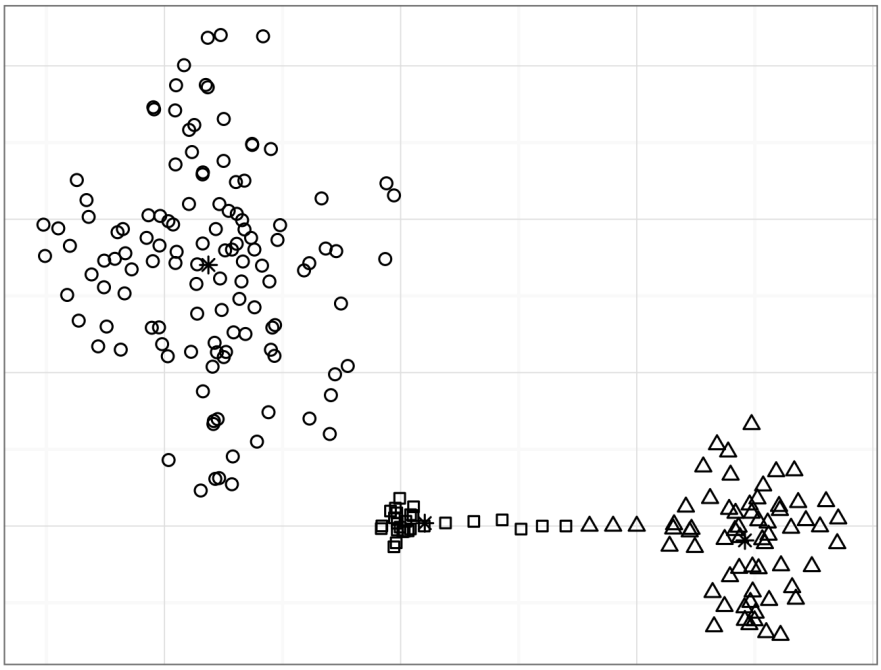
1. **[15 points]** Explain what a *ROC curve* is and how it is used.

2. **[15 points]** 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.



3. **[20 points]** Determine which of the following methods produced the clustering shown below and explain your reasoning. The centroid of each cluster is shown as an asterisk.

   - k-means clustering with $k = 3$.
   - Single linkage hierarchical clustering (dendogram cut at the level where there are 3 clusters).
   - Complete linkage hierarchical clustering (dendogram cut at the level where there are 3 clusters).



4. (a) **[5 points]** Define a high leverage point.

(b) **[5 points]** We plot a histogram of the residuals in a linear regression fit. The 10th sample has a residual that is within 2 standard deviations of the mean. Can we conclude that this point is not an outlier?

5. 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) **[10 points]** Prove that the single linkage hierarchical clustering with k clusters is the same under d and d'.

(b) **[10 points]** Prove that the complete linkage hierarchical clustering with k clusters is the same under d and d'.

6. We fit a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ to some data. Suppose we change the units of the predictors X_i, to obtain a new set of predictors $Z_i = cX_i$. Then, we fit the same data to the model: $Y = \alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_p Z_p$.

   (a) **[10 points]**
   What is the relationship between the least squares coefficients $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$ and $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$? Provide a proof.

   (b) **[10 points]** What is the relationship between the fitted values in the two models?

7. Suppose we have a classification problem with a binary response Y and a p-dimensional predictor variable $X = (X_1, \ldots, X_p)$. Logistic regression is fitted to a set of n samples. Then, logistic regression is fitted again to the same observations, where we include one additional predictor, such that:

$$X = (X_1, \ldots, X_p, X_{p+1}).$$

Explain how the training error, test error, and coefficients change in each of the following cases:

(a) $X_{p+1} = X_1 + 2X_p$.

(b) X_{p+1} is a random variable independent of Y.

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

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

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
	1	4052.181	4999.500	5403.352	5624.583
	10	10.044	7.559	6.552	5.994
d_2	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.