Your name: ____

Your SUNet ID: ____

Exam rules:

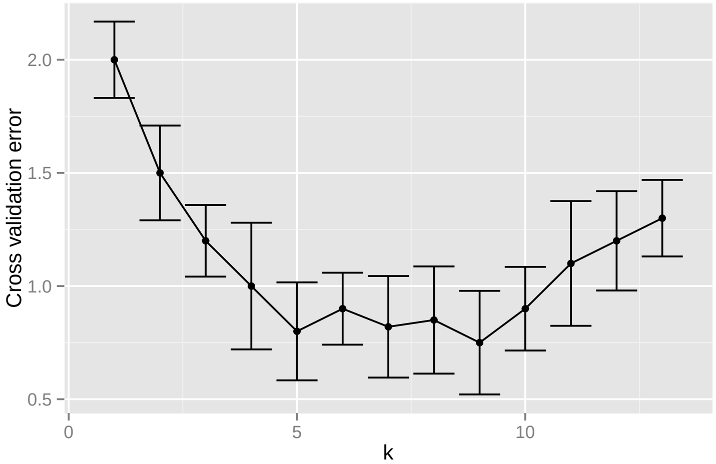
- You have until 4:00 PM July 22, 2021 to complete the exam and submit it to Gradescope.
- You are only allowed to consult your course textbooks. You are not allowed to consult other material, textbooks, computers, cell phones, the internet, or other people. If you must use a computer to type your solutions, 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.

   The Receiver Operating Characteristic (ROC) curve displays the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for a given binary classifier, under every discrimination threshold. ROC analysis provides a way to select possibly optimal models (as defined by TPR or FPR) and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.

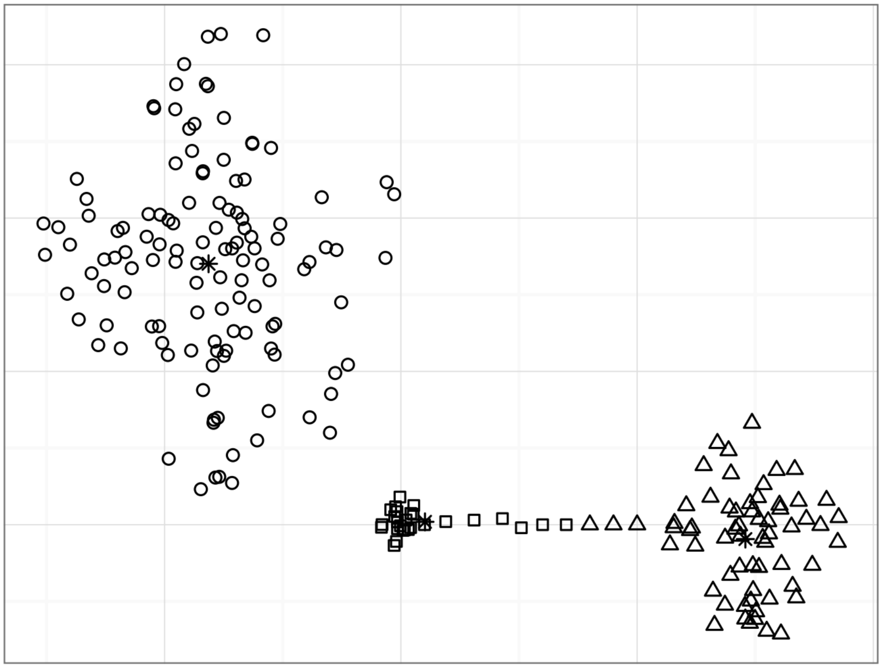
2. **[15 points]** State and explain the one standard error rule for model selection using 10-fold cross validation. Apply it to select the optimal number of nearest neighbors in the plot below, which shows the cross-validation error and one standard error intervals as a function of k.



   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 $k = 10$.

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).



The method used was complete linkage hierarchical clustering. We can eliminate 3-means clustering, because it is clear that some of the circles are closer to the centroid of the squares than to the centroid of the circles. Similarly, we can eliminate single-linkage hierarchical clustering because several circles are farther away from all other circles than the square and triangle that are closest to each other.

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

A high leverage point is a training sample which exerts an outsized influence on the fit of a linear regression because its input values are extreme. The leverage statistic measures this effect.

(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?

No. If a point has high leverage, it can have an artificially small residual while being an outlier. The studentized residuals, which are the ratio of a residual and its standard error, allow us to determine whether a high leverage point is 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'.

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) **[10 points]** 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.

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.

   In the first case, we are solving the optimization:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p)^2.$$

   In the second case, we are solving the optimization:

$$\min_{\alpha} \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 c X_1 - \cdots - \alpha_p c X_p)^2.$$

   The second problem is equivalent to the first after the change of variables $\beta_0 = \alpha_0$, $\beta_i = \alpha_i c$ for $i > 0$. This change of variables is one-to-one, i.e. every vector β maps to a unique vector α and vice versa. Therefore, the minimizer of the first problem maps to the minimizer of the second problem, or $\hat{\beta}_0 = \hat{\alpha}_0$, $\hat{\beta}_i = c\hat{\alpha}_i$ for $i > 0$.

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

   By part (a), the fitted values for the two models are equal: $\hat{y}_i = x_i^T \hat{\beta} + \hat{\beta}_0 = c x_i^T \hat{\beta}/c + \hat{\beta}_0 = z_i^T \hat{\alpha} + \hat{\alpha}_0$. One can also recall that the fitted values are a projection of the response vector y onto the column space of the predictor matrix, and $\mathbf{X}$ and $\mathbf{Z}$ have the same column space.

7. Suppose we have a classification problem with a binary response Y and a p-dimensional predictor variable $X = (X_1, \dots, X_p)$. 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:

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

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.

(a) Since the new predictor is exactly collinear with 2 of the old predictors, the coefficients β_1, β_p, and β_{p+1} are unidentifiable, as logistic regression maximizes a likelihood which only depends on a linear combination of the predictors. The predictions remain unchanged, and therefore so do the training and test errors.

(b) Since the number of samples is finite, logistic regression may assign a positive coefficient to X_{p+1} even though it is independent of the response; this will likely affect other coefficients as well. The training error can only decrease, whereas the test error will increase because the bias remains the same while variance increases.

Cheat sheet

The sample variance of $x_1, \dots, x_n$ 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.