Your name: ____

Your SUNet ID (Stanford email handle): ______________________________

Exam rules:

- You have 50 minutes to complete the exam.
- You are not allowed to consult books or notes, or to use a 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.
- If you need more space, you can attach extra sheets of paper to your solution.
- 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	
5	
Total	

1. (a) **[10 points]** Define a high leverage point.

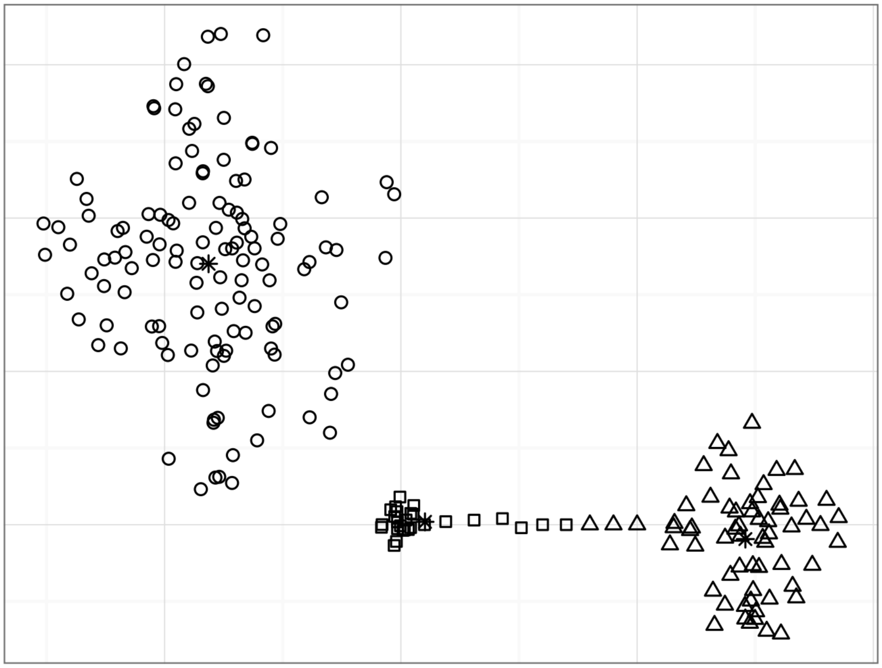
   A high leverage point is a training sample which exerts an outsized influence on the fit of a linear regression because its input values are extreme. The leverage statistic measures this effect.

   (b) **[10 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.

2. **[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.

3. The R function `ForwardSelection(y,X)` takes as arguments a vector of n outputs `y` and an n by p matrix of predictors `X`. The function performs forward stepwise selection, choosing the optimal subset of predictors through 10-fold cross validation with the 1-standard error criterion. The function outputs a linear regression fit using only the selected predictors produced by the R function `lm`.

   Consider the following R script.

```
> # Read in the input matrix
> X = read.csv(’design_matrix.csv’)
> # Read in the output vector
> y = read.csv(’output.csv’)
>
> n = length(y)
> W = rep(0,1000)
> for ( i in 1:1000 ) {
>     train = sample(n,replace=TRUE)
>     lm.fit = ForwardSelection(y[train],X[train,])
>     W[i] = length(lm.fit$coeff)-1
> }
> sd(W)
[1]  1.45
```

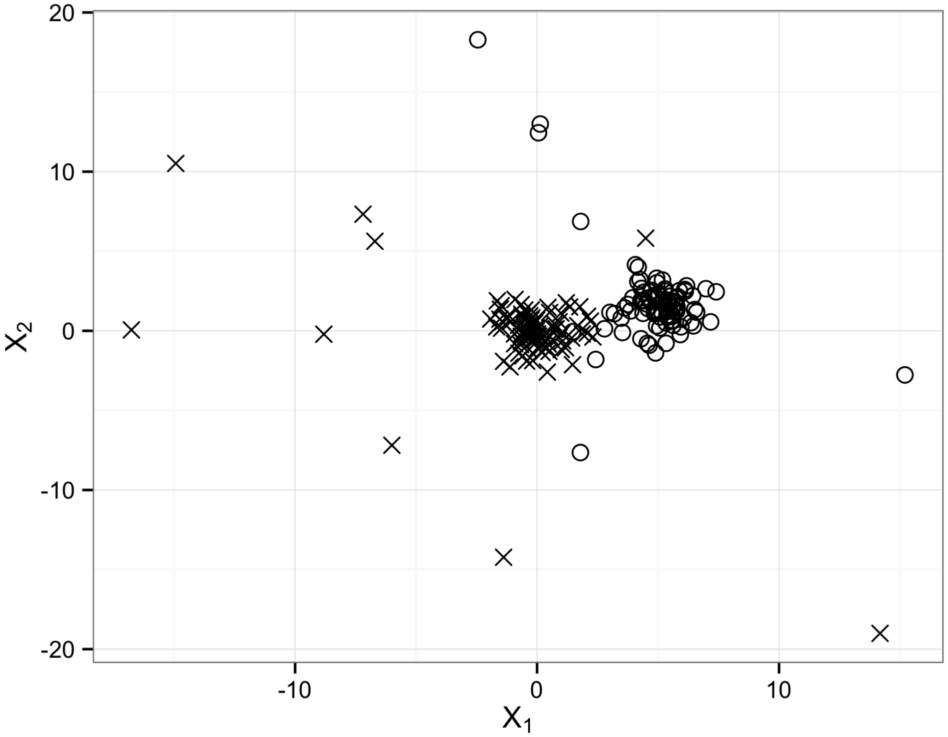
(a) **[10 points]** Name the method implemented by this script and briefly explain the logic behind it.

Bootstrapping. The method works by approximating the distribution of inputs and outputs by the empirical distribution of the training samples. This allows us to approximate the sampling distribution of diverse statistics (and their standard error) by computing the statistic on many resamplings of the training data.

(b) **[5 points]** What does the output of the script approximate?

The standard error of the number of predictors selected by `ForwardSelect`.

4. (a) **[10 points]** The figure below shows a classification dataset with a binary output (∘ or ×) and two quantitative inputs, X_1 and X_2. The standard deviation of X_2 in the class ∘ is 1. Which method would you apply to classify these data, linear discriminant analysis (LDA) or logistic regression? Explain.



From the information given, it seems several of the circular samples fall more than 10 standard deviations away from the mean. The same seems to happen in the other class. This is unlikely under a normal distribution, so the estimates of the mean and covariance used in LDA may not be robust to these outliers. Logistic regression would perform better because it learns a linear discriminant without making the assumption that samples are normal.

(b) **[10 points]** Your colleague suggests a new method called robust discriminant analysis (RDA), which is similar to LDA. The only difference is that we model the probability of the inputs given the response as a multivariate t-distribution with 2 degrees of freedom, which has density

$$P((X_1, X_2) = \mathbf{x} \mid Y = i) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}\left[1 + \frac{1}{2}(\mathbf{x} - \mu_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu_i)\right]^2},$$

where μ_i is a response-dependent vector of means of length 2, and $\boldsymbol{\Sigma}$ is a 2 by 2 covariance matrix independent of the response.

In RDA, we use maximum likelihood estimates $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\boldsymbol{\Sigma}}$ derived from the data. We approximate $P(Y = 1)$ and $P(Y = 2)$ as in LDA.

Problem: Describe the shape of the Bayes boundary of this model.

The Bayes boundary is the set of points $\mathbf{x}$ where

$$P(Y = 1 \mid \mathbf{x}) = P(Y = 2 \mid \mathbf{x}).$$

By Bayes rule,

$$P(Y = i \mid \mathbf{x}) = \frac{P(Y = i)P((X_1, X_2) = \mathbf{x} \mid Y = i)}{P(\mathbf{x})},$$

so the Bayes boundary is defined by:

$$P(Y = 1)P((X_1, X_2) = \mathbf{x} \mid Y = 1) = P(Y = 2)P((X_1, X_2) = \mathbf{x} \mid Y = 2)$$

Let $\hat{\pi}_i$ and $\hat{\pi}_2$ be the fraction of samples with response 1 and 2, respectively. Plugging our estimates for each probability in the previous equation, we obtain:

$$\frac{\hat{\pi}_1}{2\pi|\hat{\boldsymbol{\Sigma}}|^{1/2}\left[1 + \frac{1}{2}(\mathbf{x} - \hat{\mu}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu}_1)\right]^2} = \frac{\hat{\pi}_2}{2\pi|\hat{\boldsymbol{\Sigma}}|^{1/2}\left[1 + \frac{1}{2}(\mathbf{x} - \hat{\mu}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu}_2)\right]^2}$$

Eliminating identical factors on either side, and rearranging terms we get:

$$\hat{\pi}_1^{-1/2}\left[1 + \frac{1}{2}(\mathbf{x} - \hat{\mu}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu}_1)\right] = \hat{\pi}_2^{-1/2}\left[1 + \frac{1}{2}(\mathbf{x} - \hat{\mu}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu}_2)\right].$$

In general, this is a quadratic equation in $\mathbf{x}$ defining a conic section, but when $\hat{\pi}_1 = \hat{\pi}_2$, it reduces to a linear boundary.

5. [**25 points**] The reliability of a jet engine is a function f of a number of input variables $X_1, \ldots, X_{20}$ having to do with the shape and materials of its parts.

The function `SimulateReliability(x)` implemented in R takes a vector inputs x of length 20 and performs a complex computer simulation which generates a normal variate centered on $f(x)$ with variance independent of x. It is known that the reliability function f is bounded below by 0 and above by 1.

We would like to fit a linear regression model to 500 samples of the reliability at a set of 500 input vectors. The data are $(y_1, x_1), \ldots, (y_{500}, x_{500})$. We consider two methods:

- **Method 1**: Least squares multiple linear regression, $\hat{f}_1$.
- **Method 2**: Shrunk multiple linear regression, $\hat{f}_2$. This is defined as the least squares estimate multiplied by 0.8; that is $\hat{f}_2 = 0.8 \times \hat{f}_1$.

Problem: From the information given below, determine which method has a lower test MSE at an input vector $x_0 = (1, 1, \ldots, 1)$.

You are told that the true function f is roughly linear, so you can assume that least squares regression prediction $\hat{f}_1(x_0)$ is unbiased. In addition, your colleagues ran the following commands in R Studio:

```
> # Read in an input matrix of 500 rows and 20 columns
> X = read.csv('design_matrix.csv')
> # Define test input
> x0 = rep(1,20)
>
> predictions = rep(0,1000)
> for ( i in 1:1000 ) {
>     # Simulate the vector of response variables
>     y = rep(0,nrow(X))
>     for ( k in 1:nrow(X) ) {
>         y[k] = SimulateReliability(X[k,])
>     }
>     # Fit a linear model to simulated response variables
>     lm.fit = lm(y~X)
>     predictions[i] = c(1,x0) %*% lm.fit$coeff
> }
> var(predictions)
[1]  0.2
```

The test MSE at x_0 is the sum of variance, squared bias, and irreducible error. The irreducible error is the same for both estimators. The bias of $\hat{f}_1(x_0)$ is 0, while the bias of $\hat{f}_2(x_0)$ is:

$$\begin{aligned} E(\hat{f}_2(x_0)) - f(x_0) &= E(0.8 \times \hat{f}_1(x_0)) - f(x_0) \\ &= 0.8(E(\hat{f}_1(x_0)) - f(x_0)) - 0.2f(x_0) = -0.2f(x_0). \end{aligned}$$

Since the function f takes values inside $[0,1]$, we conclude that the squared bias is at most $0.2^2 = 0.04$.
The R code provided approximates the variance of $\hat{f}_1(x_0)$, which is around 0.2. On the other hand,

$$\text{Var}(\hat{f}_2(x_0)) = \text{Var}(0.8\hat{f}_1(x_0)) \approx 0.8^2 \times 0.2 = 0.64 \times 0.2 = 0.128.$$

So, the drop in variance due to using the shrunk estimator is higher than the increase in square bias, and we can conclude that Method 2 has a lower test MSE.

Cheat Sheet

Tail probabilities of the standard normal distribution

z	1	2	3	4	5
$P(Z > z)$	1.586553e-01	2.275013e-02	1.349898e-03	3.167124e-05	2.866516e-07

Bayes rule

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Linear equation in two variables $x = (x_1, x_2)^T$

$$Mx + c = 0$$

where M is a 1 by 2 matrix and c is a number.

Quadratic equation in two variables $x = (x_1, x_2)^T$ **(conic section)**

$$x^T M_1 x + M_2 x + c = 0$$

where M_1 is a 2 by 2 matrix, M_2 a 1 by 2 matrix, and c is a number.