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.

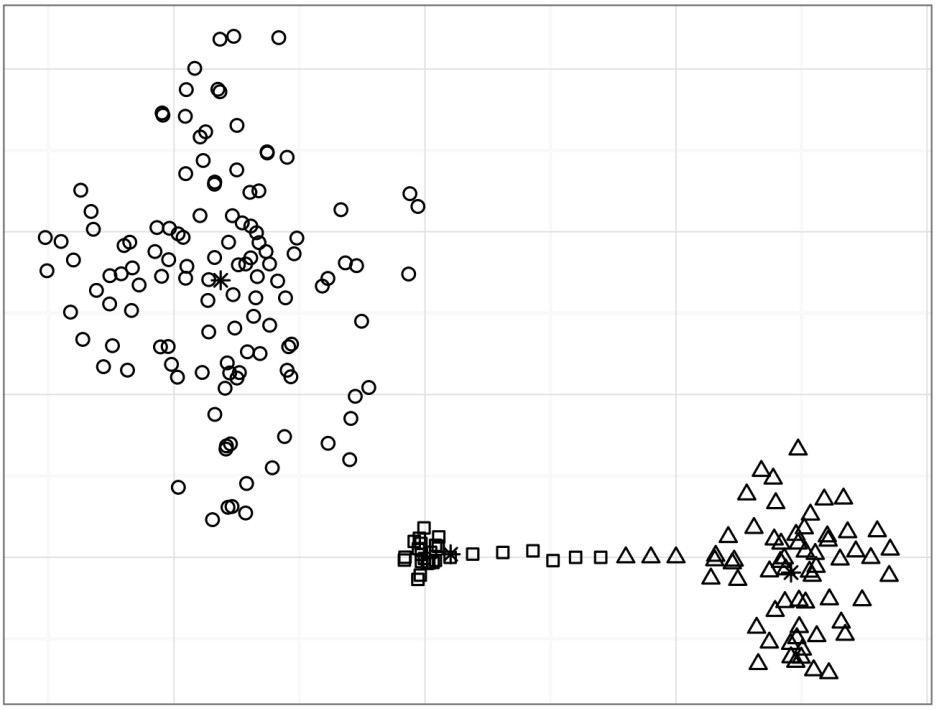
Problem	Points
1	
2	
3	
4	
5	
Total	

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

   (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?

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).



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`.

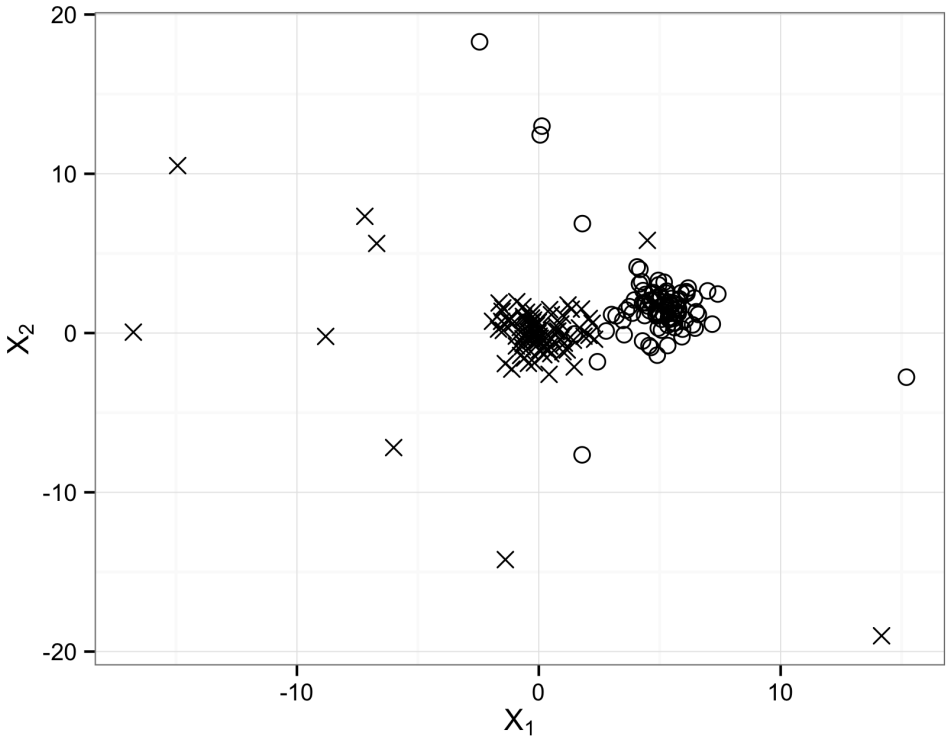
   Consider the following R script.

```
> # 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.

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

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



(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

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

where μ_i is a response-dependent vector of means of length 2, and $\mathbf{\Sigma}$ 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 $\hat{\mathbf{\Sigma}}$ 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.

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
```

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.