MAT 191 Fall 2015
Midterm Exam

Name ____

ID __________________

Problem	1	2	3	4
Score				

Working on the exam:

- the answers must represent your own work only;
- use of your class notes (including corrected and annotated homework assignments), of material posted on the CCLE website for this course, and of the textbook
  - *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani,

  freely available at `http://www-bcf.usc.edu/~gareth/ISL/`.
- the above are the only materials you can use; in particular, no other books, no internet resources, no exchanges with other people (with the exception of myself, in case you need some clarification) are allowed

Returning the exam:

- hand in the exam in class **on Monday, November 16**. If you prefer to hand it in earlier, you can always slide the exam under my office door, in MS 7310.
- No late submissions or partial submissions will be accepted whatsoever.

Signed statement: I confirm that I have followed all the rules for this written take-home examination, and that this represents my own work in accordance with University regulations.

____

____

____

____

____

Explain all your answers, and PLEASE write clearly and neatly.

Problem 1 (10p)

(a)(5p) Let X be a random variables distributed uniformly $\sim$ Unif($[-1,1]$) and let $Y = X^2$. Are X, Y correlated? Are they independent?

(b)(5p) Compute the characteristic function of the following distribution

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{if } x \notin (a,b) \end{cases}$$

Problem 2 (10p)

Assume that $x_1, x_2, \ldots, x_n$ are independent identically distributed instances of a p-dimensional random variable X with mean μ (of size $p \times 1$) and covariance C (of size $p \times p$). Given that the sample mean is defined as

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k$$

show the following result: $\mathbf{Var}[\mu_n] = \frac{1}{n} C$.

Problem 3 (10p) Problem 9 from the textbook, Chapter 3, page 122

Problem 4 (10p)

(a)(5p) When we discussed multidimensional scaling in class, we made the following claim. If $x_1, \ldots, x_n \in \mathbb{R}^p$, $D_{ij} = ||x_i - x_j||_2^2$, $s_i = \sum_{j=1}^{n} D_{ij}$, and $s = \sum_{i=1}^{n} s_i$ then

$$D_{ij} - \frac{1}{n} s_i - \frac{1}{n} s_j + \frac{1}{n^2} s = -2 x_i^T x_j$$

Prove this claim.

(b)(5p) Download the associated data set from CCLE, named "**midtermDistanceMatrix.data**" and perform multidimensional scaling on it. After loading the file, with the command *load(file = 'midtermDistanceMatrix.data')*, the matrix of distances is called DIST. More precisely, you are asked to:

(i) Build the matrix B we constructed in class, and explore its spectrum. Show a barplot of the top 10 eigenvalues. What do you observe, and what can you conclude about the space in which the point $x_1, \ldots, x_n$ live?

(ii) Compute a 2-dimensional embedding, by using the spectral decomposition of B, as shown in class (You may want to reflect one of the axes in your plot, to make the end result more visually appealing). Include the embedding in your submission, as well as the code used for the entire part (b).