

Assignment 5
Submission deadline: Friday Nov 4, 2016 by 8 PM
Submission format: upload document in Canvas

1. **(20 points) BONUS: complete the laptop cores survey on Canvas by Sunday, Oct 30th 2016.**
2. **(40 points) Principal Components Analysis.**
 - (a) Load the **USArrests** dataset in R, and describe the dataset in your own words, in 2-3 lines.
 - (b) Calculate the mean and standard deviation of the four variables.
 - (c) Perform principal components analysis of the dataset using R. How many principal components are there? For each principal component, report the standard deviation and the proportion of variance explained.
 - (d) How many principal components would you need to explain at least (i) 60% of the total variance? (ii) 75% of the total variance? (iii) 90% of the total variance?
3. **(60 points) Hierarchical and k-means clustering.**
 - (a) Load the **NCI60** dataset from the **ISLR** package in R, and describe the dataset in your own words, in 2-3 lines.
 - (b) Using Euclidean distance as the dissimilarity measure, perform hierarchical clustering on the data, with (i) Complete Linkage, (ii) Average Linkage, and (iii) Single Linkage.
 - (c) For all three methods in (b), cut the hierarchical clustering tree at 4 clusters, and report the two-way table of actual cancer types (**NCI60\$labs**) and clusters. Are there any differences between the tables from different methods?
 - (d) Perform k-means clustering of the data with k=4 clusters. Report the two-way table of actual cancer types (**NCI60\$labs**) and clusters.

Assignment instructions:

1. **Honor code:** The Virginia Tech honor pledge for assignments is as follows:
"I have neither given nor received unauthorized assistance on this assignment."

The pledge is to be written out on all graded assignments at the university and signed by the student. Type up your name to sign.
2. Submit your assignment as a document (word, pdf or similar) to Canvas, clearly marked with student's name and assignment number, eg. Sengupta_Srijan_HW5.pdf. Your submission should include R code and answers to problems.

3. Late assignments will not be accepted. Check Canvas regularly for assignments and submission dates.
4. You are free to discuss assignment problems with your classmates, but submitted work (answers and codes) **must** be your own work. Students are not allowed to copy computer codes or answers from each other, and must write their own codes and answers.