Written homework 5

Math 187: Introduction to Applied Linear Algebra

Due in class: Wednesday, October 2

1. 3.11 on page 65

2. The U.S. healthcare system uses patient satisfaction surveys to determine whether the amount patients pay and the quality of care they receive is justified. While we don't have access to that survey data, we do have typical data of this type.

   The dataset `PatientSatisfaction.csv` contains responses from 46 patients who were given a satisfaction survey as they were released from the hospital. The columns of the dataset list the variables measured by the survey: satisfaction (on a 0-100 scale with larger values indicating more satisfaction), anxiety (higher numbers indicating more anxiety), severity (larger numbers meaning more severe illness), and age. Each row corresponds to a unique patient's response.

   Note that in order to carry out computations in R, you may find the following built-in functions useful:

Math	R
$\sqrt{}$	`sqrt()`
$\sum$	`sum()`
average	`mean()`

   (a) Use R to calculate the average values for each variable. Record those averages on paper.

   (b) Use R to calculate the de-meaned vectors for each variable. (These will be long lists of numbers! You do not need to record these on paper.)

   (c) Use the de-meaned vectors to calculate in R the standard deviations for each variable. Record those standard deviations on paper.

   (d) At some point, we'd like to group the patients into clusters. You've already thought about this in the previous homework problem. Use R to calculate the Euclidean distance between patients 1 and 2. How far apart are they? Repeat this for patients 1 and 3, patients 1 and 5, and again for patients 1 and 10.

   (e) From your previous calculations, which patients would you say are most similar? least similar?

   (f) Use R to create standardized variables (i.e., z-scores).

(g) What does the standardized value for patient 9's patient satisfaction score tell you about this patient's satisfaction level?

(h) Using the standardized variables, use R to calculate the Euclidean distance between patients 1 and 2. How far apart are they? Repeat this for patients 1 and 3, patients 1 and 5, and again for patients 1 and 10.

(i) From these calculations, which patients would you now say are most similar? least similar?

(j) Imagine we created a vector for each patient that consisted only of the z-scores for *anxiety* and *severity* variables.

   i. What kind of angle would you expect between two such vectors if one patient had above average anxiety and above average severity, and the other patient had below average anxiety and below average severity?

   ii. What about if the second patient also had above average anxiety and above average severity?

   iii. What could you say if the cosine similarity between the two vectors was 0? Explain. *Note that cosine similarity is the cosine of the angle between two vectors.*

(k) What did you notice in your analysis of this data? What do you still have questions about?