Homework 6 (STAT 5860)

Your Name Here

Due by 11:59 am, Apr. 20, 2018

Insturctions:

1. Download the `Homework6.Rmd` file from the course Elearning.
2. Open `Homework6.Rmd` in RStudio.
3. Replace the *"Your Name Here"* text in the `author` with your own name.
4. Write your answer to each problem by editing `Homework6.Rmd`.
5. After you finish all the problems, click `Knit to PDF` to create a pdf file. Upload your pdf file to Homework 6 Dropbox in the course Elearning.

Set the seed number

```
set.seed(413)
```

Problem 1. Use the `solveLP()` function in `linprog` package to solve the following Linear Prgoramming problem.

$$\text{Minimize} \quad 4x + 2y + 9z \quad \text{subject to}$$

$$2x + y + z \leq 2$$

$$-x + y - 3z \geq -3$$

$$x \geq 0, y \geq 0, z \geq 0$$

```
# Write R code
```

Problem 2. In this problem, we will peform simulation studies to compare the hard-clustering results between EM algorithm and K-means clustering. To assess the clustering performance, we will use the Rand Index (Rand 1971) which measures the similarity between two clustering otucomes. Since this is a simulation study, we know the true cluster id for each observation. Hence you can use the Rand Index to measure the similarity between true cluster id and clustering outcome and this will tell you how good is your method. The Rand index has a value between 0 and 1, with 0 indicating that the two clusterings do not agree on any pair of observations and 1 indicating that the clusterings are exactly the same. You can use `rand.index()` function in `fossil` pacakge to calculate the Rand Index.

(a) Generate a random sample of size 1000 from the following mixture of normal distribution. Also, draw a histogram to see how the distribution looks like.

$$0.5 \times N(-1, 0.5^2) + 0.5 \times N(1, 1^2)$$

```
# Write R code
```

(b) Use EM algorithm to obtain posterior probabilities of memebrship for observations. Now we can obtain hard-clustering result by assiging observation to the cluster where the posterior porbablity value is largest. Use true cluster id and hard-clustering result to calcualte the Rand Index. (Hint: you can use `normalmixEM()` function in `mixtools` package to use EM algorithm.)

```
# Write R code
```

(c) Use K-means algorithm to obtain hard-clustering result. Use true cluster id and hard-clustering reuslt to calculate the Rand Index. (Hint: you can use `kmeans()` function in R to use K-means algorithm.)

```
# Write R code
```

(d) Compare the two Rand Index from (b) and (c) and make comments on the result.

(e) Genearte a random sample of size 1000 from the following mixture of normal distribution and draw a histogram to see how the distribution looks like. Now repeat (b) - (d). Also, make comments on how the distribution change affects the clustering performance of both EM algorithm and K-means clusteirng.

$$0.5 \times N(-1, 0.5^2) + 0.5 \times N(1, 0.5^2)$$

```
# Write R code
```