Stat 437 Project 1

Your Name (Your student ID)

General rule and information

You must show your work in order to get points. Please prepare your report according to the rubrics on projects that are given in the syllabus. In particular, please note that your need to submit codes that would have been used for your data analysis. Your report can be in .doc, .docx, .html or .pdf format.

The project will assess your skills in K-means clustering,Hierarchical clustering, Nearest-neighbor classifier, and discriminant analysis for classification, for which visualization techniques you have learnt will be used to illustrate your findings.

Data set and its description

Please download the data set "TCGA-PANCAN-HiSeq-801x20531.tar.gz" from the website https://archive.ics.uci.edu/ml/machine-learning-databases/00401/. A brief description of the data set is given at https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq.

You need to decompress the data file since it is a .tar.gz file. Once uncompressed, the data files are "labels.csv" that contains the cancer type for each sample, and "data.csv" that contains the "gene expression profile" (i.e., expression measurements of a set of genes) for each sample. Here each sample is for a subject and is stored in a row of "data.csv". In fact, the data set contains the gene expression profiles for 801 subjects, each with a cancer type, where each gene expression profile contains the gene expressions for the same set of 20531 genes. The cancer types are: "BRCA", "KIRC", "COAD", "LUAD" and "PRAD". In both files "labels.csv" and "data.csv", each row name records which sample a label or observation is for.

Task A. Clustering

For this task, you need to apply k-means and hierarchical clustering to cluster observations into their associated cancer types, and report your findings scientifically and professionally. Your laptop may not have sufficient computational power to implement k-means and hierarchical clustering on the whole data set, and genes whose expressions are zero for most of the subjects may not be so informative of a cancer type.

Please use `set.seed(123)` for random sampling via the command `sample`, random initialization of `kmeans`, implementing the gap statistic, and any other process where artificial randomization is needed.

(**Task A1**) Complete the following data processing steps:

- Filter out genes (from "data.csv") whose expressions are zero for at least 300 subjects, and save the filtered data as R object "gexp2".

- Use the command `sample` to randomly select 1000 genes and their expressions from "gexp2", and save the resulting data as R object "gexp3".
- Use the command `sample` to randomly select 30 samples and their labels from the file "labels.csv", and save them as R object "labels1". For these samples, select the corresponding samples from "gexp3" and save them as R object "gexpProj1".
- Use the command `scale` to standard the gene expressions for each gene in "gexpProj1", so that they have sample standard deviation 1. Save the standardized data as R object "stdgexpProj1".

(**Task A2**)

(**Part 1 of Task A2**) Randomly pick 50 genes and their expressions from "stdgexpProj1", and do the following to these expressions: apply the "gap statistic" to estimate the number of clusters, apply K-means clustering with the estimated number of clusters given by the gap statistic, visualize the classification results using techniques given by "LectureNotes3_notes.pdf.pdf", and provide a summary on classification errors. You may use the command `table` and "labels1" to obtain classification errors. Note that the cluster numbering given by `kmeans` will usually be coded as follows:

```
#   Class  label
#     PRAD  5
#     LUAD  4
#     BRCA  1
#     KIRC  3
#     COAD  2
```

When you apply `clusGap`, please use arguments `K.max=10, B=200,iter.max=100`, and when you use `kmeans`, please use arguments `iter.max = 100, nstart=25, algorithm = c("Hartigan-Wong")`.

(**Part 2 of of Task A2**) Upon implementing `kmeans` with k as the number of clusters, we will obtain the "total within-cluster sum of squares" $W(k)$ from the output `tot.withinss` of `kmeans`. If we try a sequence of $k = 1, 2, 3, ..., 10$, then we get $W(k)$ for each k between 1 and 10. Let us look at the difference $\Delta_k = W(k) - W(k+1)$ for k ranging from 1 to 9. The K^* for which

$$\{\Delta_k : k < K^*\} \gg \{\Delta_k : k \geq K^*\}$$

is an estimate of the true number K of clusters in the data, where $\gg$ means "much larger". Apply this method to obtain an estimate of K for the data you created in **Part 1**, and provide a plot of $W(k)$ against k for each k between 1 and 10. Compare this estimate with the estimate obtained in **Part 1** given by the gap statistic, comment on the accuracy of the two estimates, and explain why they are different.

(**Part 3 of of Task A2**) Randomly pick 250 genes and their expressions from "stdgexpProj1", and for these expressions, do the analysis in **Part 1** and **Part 2**. Report your findings, compare your findings with those from **Part 1** and **Part 2**; if there are differences between these findings, explain why. Regard using more genes as using more features, does using more features necessarily give more accurate clutering or classification results?

(**Task A3**) Randomly pick 250 genes and their expressions from "stdgexpProj1", and for these expressions, do the following: respectively apply hierarchical clustering with average linkage, single

linkage, and complete linkage to cluster subjects into groups, and create a dendrogram. For the dendrogram obtained from average linkage, find the height at which cutting the dendrogram gives the same number of groups in "labels1", and comment on the clustering results obtained at this height by comparing them to the truth contained in "labels1".

Task B. Classification

For this task, we will use the same data set you would have downloaded. Please use `set.seed(123)` for random sampling via the command `sample` and any other process where artificial randomization is needed.

(**Task B1**) After you obtain "labels.csv" and "data.csv", do the following:

- Filter out genes (from "data.csv") whose expressions are zero for at least 300 subjects, and save the filtered data as R object "gexp2".
- Use the command `sample` to randomly select 1000 genes and their expressions from "gexp2", and save the resulting data as R object "gexp3".
- Pick the samples from "labels.csv" that are for cancer type "LUAD" or "BRCA", and save them as object "labels2". For these samples, pick the corresponding gene expressions from "gexp3" and save them as object "stdgexp2"

(**Taks B2**) The assumptions of linear or quadratic discriminant analysis requires that each observation follows a Gaussian distribution given the class or group membership of the observation, and that each observation follows a Gaussian mixture model. In our settings here, each observation (as a row) within a group would follow a Gaussian with dimensionality equal to the number of genes (i.e., number of entries of the row). So, the more genes whose expressions we use for classification, the higher the dimension of these Gaussian distributions. Nonetheless, you need to check if the Gaussian mixture assumption is satisfied. Note that we only consider two classes "LUAD" and "BRCA", for which the corresponding Gaussian mixture has 2 components and hence has 2 bumps when its density is plotted.

Do the following and report your findings on classification:

- Randomly pick 3 genes and their expressions from "stdgexp2", and save them as object "stdgexp2a".
- Randomly pick 60% of samples from "stdgexp2a", use them as the training set, and use the rest as the test set. You can round down the number of samples in the training set by the command `floor` if it is not an integer.

Build a quadratic discriminant analysis model using the training set, and apply the obtained model to the test set to classify each of its observations. You should code "BRCA" as 0 and "LUAD" as 1. If for an observation the posterior probability of being "BRCA" is predicted by the model to be greater than 0.5, the observation is classified as "BRCA". Report via a 2-by-2 table on the classification errors. Note that the predicted posterior probability given by `qda` is for an observation to belong to class "BRCA".

Before building a quadratic discriminant analysis model, you need to check for highly correlated gene expressions, i.e., you need to check the sample correlations between each pair of columns of

the training set. If there are highly correlated gene expressions, the estimated covariance matrix can be close to to being singular, leading to unstable inference. You can remove a column from two columns when their contained expressions have sample correlation greater than 0.9 in absolute value.

(**Taks B3**) Do the following:

- Randomly pick 100 genes and their expressions from "stdgexp2", and save them as object "stdgexp2b".
- Randomly pick 75% of samples from "stdgexp2b", use them as the training set, and use the rest as the test set. You can round down the number of samples in the training set by the command `floor` if it is not an integer.

Then apply quadratic discriminant analysis by following the requirements given in **Taks B2**. Compare classification results you find here with those found in **Taks B2**, and explain on any difference you find between the classification results.

(**Taks B4**) Do the following:

- Randomly pick 100 genes and their expressions from "stdgexp2", and save them as object "stdgexp2b".
- Randomly pick 75% of samples from "stdgexp2b", use them as the training set, and use the rest as the test set. You can round down the number of samples in the training set by the command `floor` if it is not an integer.

Then apply k-nearest-neighbor (k-NN) method with neighborhood size k=3 to the test data to classify each observation in the test set into one of the cancer types. Here, for an observation, if the average of being cancer type "BRCA" is predicted by k-NN to be greater than 0.5, then the observation is classified as being "BRCA". Report via a 2-by-2 table on the classification errors. Compare and comment on the classification results obtained here to those obtain in **Taks B3**. If there is any difference between the classification results, explain why.