Clustering Analysis
Babelomics 5.0
Exercises

Cankut ÇUBUK
March 9th, 2017

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Babelomics 5.0

http://babelomics.bioinfo.cipf.es/

http://courses.babelomics.org/

Exercises

1) Go to http://bioinfo.cipf.es/gda17/doku.php/program

2) Download GDA17_TCGA_265_mod_gene_BRCA_subtype_HER_Basal_Normal.txt (Prediction_clustering_exercises.zip)

   Data description: RNA-Seq data of 30 Breast Invasive Carcinoma (BRCA) samples taken from The Cancer Genome Atlas (TCGA) data portal. Contains 10 normal samples, 20 tumor samples with 2 subtypes (Basal-like and Her2-enriched).

3) Upload your file to Babelomics 5.0.

   Go to section Expression>Clustering

4) Cluster samples with given paramaters.

   UPGMA + Euclidean (square)

   UPGMA + Correlation coeff. (Spearman)

   Which distance parameter is better for proper clustering?

Exercises

5) Repat the analysis using the same distance parameters and SOTA method.

SOTA + Euclidean (square)

SOTA + Correlation coeff. (Spearman)

Do the results change based on the method or the distance parameter?

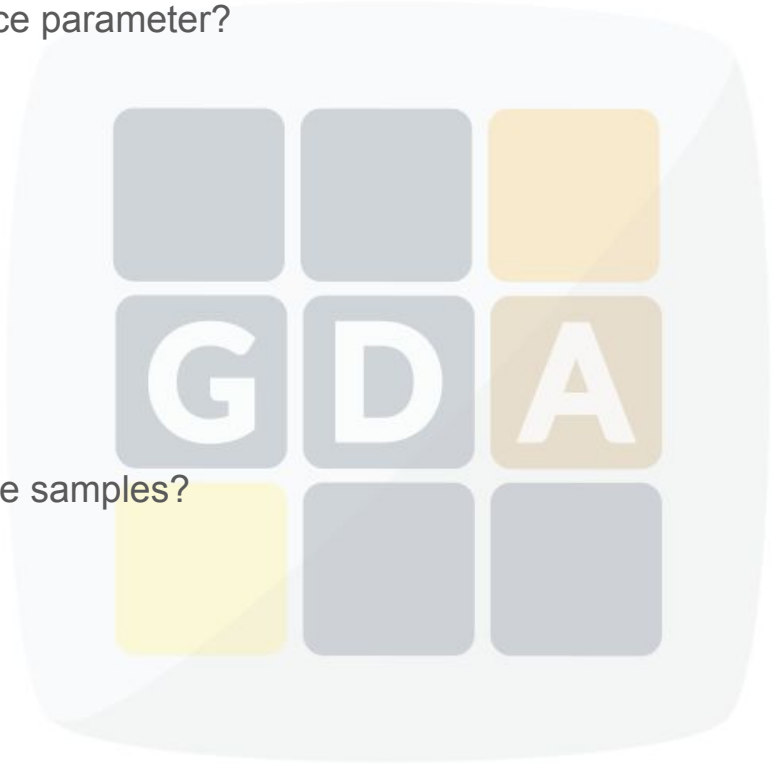
6) Try to cluster your samples with K-means.

Set k-value 6 and use Correlation coeff. (Spearman)

Check the results of K-means.

Are the results acceptable?

Is the dendrogram representing any hierarchy between the samples?



Exercises

7) Repeat the step 6 with k-value 3.

Did your result same as previous one?

8)) Try to cluster your samples with K-means.

Set k-value 2 and use Correlation coeff. (Spearman).

Can we say that K-means is good to distinguish tumor from normal?

Supervised Classification Analysis
Babelomics 5.0
Exercises

Cankut ÇUBUK
March 9th, 2017

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Babelomics

http://babelomics.bioinfo.cipf.es/

http://courses.babelomics.org/

Exercises

1) Go to http://bioinfo.cipf.es/gda17/doku.php/program

2) Download GDA17_TCGA_265_mod_gene_LUSC_train.txt (Prediction_clustering_exercises.zip)

Data description: RNA-Seq data of Lung squamous cell carcinoma (LUSC) samples taken from The Cancer Genome Atlas (TCGA) data portal.

Contains 11 Normal and 150 Tumor samples.

NOTE: It is always recommended to use balanced sample size to avoid biased training.

3) Download GDA17_TCGA_265_mod_gene_LUSC_test.txt (Prediction_clustering_exercises.zip)

Data description: RNA-Seq data of Lung squamous cell carcinoma (LUSC) samples taken from The Cancer Genome Atlas (TCGA) data portal.

Contains 6 Normal and 75 Tumor samples.

4) Upload your files to Babelomics 5.0.

   Go to section Expression>Class Prediction

Exercises

5) Select SVM, KNN and Random Forest

   Select Leave-one-out for error estimation

   Select Correlation-based Feature Selection (CFS)

6) Download test_result.txt

Which supervised classification method(s) works better?

How many genes were used for the prediction?

Are the selected genes same for all methods?