← Go Back to Data Analysis & Visualization

Course Content

Clarification Note - Genomic Data Clustering

The frequency table displayed is the following:

	aaa	aag	aat	aac	aga	agg	agt	agc	ata	atg	...	cgt	cgc	cta	ctg	ctt	ctc	cca	ccg	cct	ccc
0	0.0	1.0	0.0	3.0	0.0	2.0	0.0	2.0	2.0	0.0	...	2.0	2.0	2.0	1.0	1.0	4.0	4.0	4.0	1.0	1.0
1	1.0	0.0	0.0	3.0	0.0	1.0	0.0	0.0	5.0	0.0	...	2.0	4.0	0.0	2.0	4.0	2.0	1.0	1.0	0.0	1.0

In the video, at 3:20, the instructor says “the PCA table is going to have one row for every column of the frequency table".

```
    pca = PCA(n_components = 2)

    pCompTables = {}

    for i in range(1,5):
        pca.fit(normFreqTables[i])
        pComponents = pca.transform(normFreqTables[i])
        # for each word size, we store the result of the PCA in a table containing only the 2 principipal components
        pCompTables[i] = pd.DataFrame(pComponents[:, [0,1]], columns = ['pc1', 'pc2'])
        print('Explained variance for ' + str(i) + ' letters: ' + str(pca.explained_variance_ratio_.sum()))

    print(pCompTables[2].head())
[8]
...  Explained variance for 1 letters: 0.7489363490534278
     Explained variance for 2 letters: 0.22774966356188592
     Explained variance for 3 letters: 0.31670201938180154
     Explained variance for 4 letters: 0.029334525098403036
            pc1       pc2
     0 -0.949207 -0.396447
     1 -0.100967 -0.873554
     2  1.198262  0.366900
     3  0.579445  1.174832
     4  0.102275  1.788382
```

Here, this refers to the fact that for every entry/row in the frequency table we calculate it. But, the other part of the explanation may be a little confusing.

At first, we scale the ‘freqTables’ into ‘normFreqTables’. If you print 'normFreqTables[1] or normFreqTables[2]', you will be able to see the kind of data we are fitting the PCA algorithm into:

```
normFreqTables[1]
✓ 0.1s
```

	a	g	t	c
0	0.589191	0.247731	-0.046647	-0.864764

```
normFreqTables[2]
```

✓ 0.1s Python

	aa	ag	at	ac	ga	gg	gt	gc	ta	tg	tt	tc	ca	cg	ct	
0	0.678394	3.017473	0.366334	-0.364310	-1.939741	1.320479	1.380430	-0.624884	0.896849	-0.830630	0.238082	-0.044287	-0.705184	-2.260329	-1.603402	1.785

We fit PCA to the above data and then, the 'pCompTables' shows the value of components for each entry in the corresponding normFreqTables.

So that means pCompTables[1] is the Principal Component table that we get from normFreqTables[1], and similarly, we calculate pCompTables[2] using normFreqTables[2].

Previous Next