My Extension for AAS_CH5 & DM4.8

Alexander Spivey

What dataset are we using?

1. Linkage

2. Covtype.data

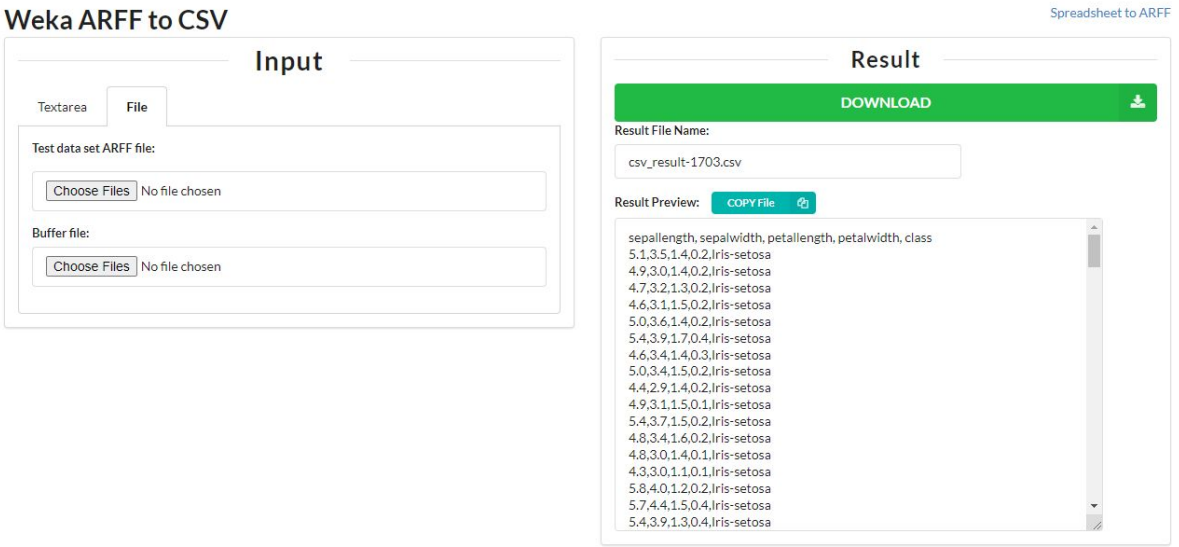
3. iris.arff

Linkage and Covtype

```
scala> parsed.printSchema()
root
 |-- id_1: string (nullable = true)
 |-- id_2: integer (nullable = true)
 |-- cmp_fname_c1: double (nullable = true)
 |-- cmp_fname_c2: double (nullable = true)
 |-- cmp_lname_c1: double (nullable = true)
 |-- cmp_lname_c2: double (nullable = true)
 |-- cmp_sex: integer (nullable = true)
 |-- cmp_bd: integer (nullable = true)
 |-- cmp_bm: integer (nullable = true)
 |-- cmp_by: integer (nullable = true)
 |-- cmp_plz: integer (nullable = true)
 |-- is_match: boolean (nullable = true)


scala> :load me1.scala
Loading me1.scala...
import org.apache.spark.ml.{PipelineModel, Pipeline}
import org.apache.spark.ml.clustering.{KMeans, KMeansModel}
import org.apache.spark.ml.feature.{OneHotEncoder, VectorAssembler, StringIndexer, Stand
ardScaler}
import org.apache.spark.ml.linalg.{Vector, Vectors}
import org.apache.spark.sql.{DataFrame, SparkSession}
import scala.util.Random
parsed: org.apache.spark.sql.DataFrame = [id_1: string, id_2: int ... 10 more fields]
numericData: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cmp_fname_c1: dou
ble, cmp_fname_c2: double ... 8 more fields]
assembler: org.apache.spark.ml.feature.VectorAssembler = vecAssembler_b77ad4bd49d6
kmeans: org.apache.spark.ml.clustering.KMeans = kmeans_6c2a3308533a
pipeline: org.apache.spark.ml.Pipeline = pipeline_c02c930ee3aa
[Stage 11:>                                                         (0 + 8) / 10]21/09/25
 17:25:23 WARN BlockManager: Putting block rdd_52_1 failed due to an exception
21/09/25 17:25:23 WARN BlockManager: Block rdd_52_1 could not be removed as it was not f
ound on disk or in memory
21/09/25 17:25:23 WARN BlockManager: Putting block rdd_52_3 failed due to an exception
21/09/25 17:25:23 WARN BlockManager: Block rdd_52_3 could not be removed as it was not f
ound on disk or in memory
21/09/25 17:25:23 WARN BlockManager: Putting block rdd_52_4 failed due to an exception
21/09/25 17:25:23 WARN BlockManager: Block rdd_52_4 could not be removed as it was not f
ound on disk or in memory
21/09/25 17:25:23 WARN BlockManager: Putting block rdd_52_6 failed due to an exception
21/09/25 17:25:23 WARN BlockManager: Block rdd_52_6 could not be removed as it was not f
ound on disk or in memory
21/09/25 17:25:23 WARN BlockManager: Putting block rdd_52_2 failed due to an exception
21/09/25 17:25:23 WARN BlockManager: Block rdd_52_2 could not be removed as it was not f
ound on disk or in memory
21/09/25 17:25:23 ERROR Executor: Exception in task 3.0 in stage 11.0 (TID 50)
org.apache.spark.SparkException: Failed to execute user defined function($anonfun$3: (st
ruct<cmp_fname_c1:double,cmp_fname_c2:double,cmp_lname_c1:double,cmp_lname_c2:double,cmp
_sex_double_vecAssembler_b77ad4bd49d6:double,cmp_bd_double_vecAssembler_b77ad4bd49d6:dou
ble,cmp_bm_double_vecAssembler_b77ad4bd49d6:double,cmp_by_double_vecAssembler_b77ad4bd49
d6:double,cmp_plz_double_vecAssembler_b77ad4bd49d6:double>) => vector)
```

Iris.arff to CSV

https://pulipulichen.github.io/jieba-js/weka/arff2csv/



Or just do it by hand

Initial Runs

cluster	class	count
0	Iris-setosa	50
0	Iris-versicolor	3
1	Iris-virginica	50
1	Iris-versicolor	47

```
clusteringScore0: (data: org.apache.spark.sql.DataFrame, k: Int)Double
(2,1.0157913765156006)
(3,0.5262722761743098)
(4,0.38236679365079446)
(5,0.33172117826617714)
(6,0.259539753664487)
(7,0.2315311253561297)
(8,0.21983692826765452)
(9,0.2034134334728438)
(10,0.17869285992760023)
```

```
clusteringScore1: (data: org.apache.spark.sql.DataFrame, k: Int)Double
(2,1.0157913765156006)
(3,0.5262722761743098)
(4,0.47565482345522514)
(5,0.3118806666666675)
(6,0.29927116817065413)
(7,0.2315311253561297)
(8,0.22724368828681013)
(9,0.22565139682540103)
(10,0.1856260975950651)
clusteringScore2: (data: org.apache.spark.sql.DataFrame, k: Int)Double
(2,1.4816030602123282)
(3,0.9410882192233426)
(4,0.7628156641257224)
(5,0.681093993781662)
(6,0.5497218951783667)
(7,0.48113897137396633)
(8,0.4616568214947095)
(9,0.38923413476602076)
(10,0.3536573744307816)
```

Scaled vs Non-scaled

```
clusteringScore4: (data: org.apache.spark.sql.DataFrame, k: Int)Double
(2,0.4620981203732969)
(3,0.37448781399842235)
(4,0.3941260666425602)
(5,0.325552825562554)
(6,0.253110107336976)
(7,0.22338263109504614)
(8,0.2507263834776569)
(9,0.1245837434615639)
(10,0.19430331850536775)
```

```
clusteringScore5: (data: org.apache.spark.sql.DataFrame, k: Int)Double
(2,0.524782410959326)
(3,0.2895730091214502)
(4,0.2366309466813321)
(5,0.21058876834680107)
(6,0.19348437504568017)
(7,0.08282821774921538)
(8,0.09290223403419938)
(9,0.08086951653803698)
(10,0.14721739767228123)
```

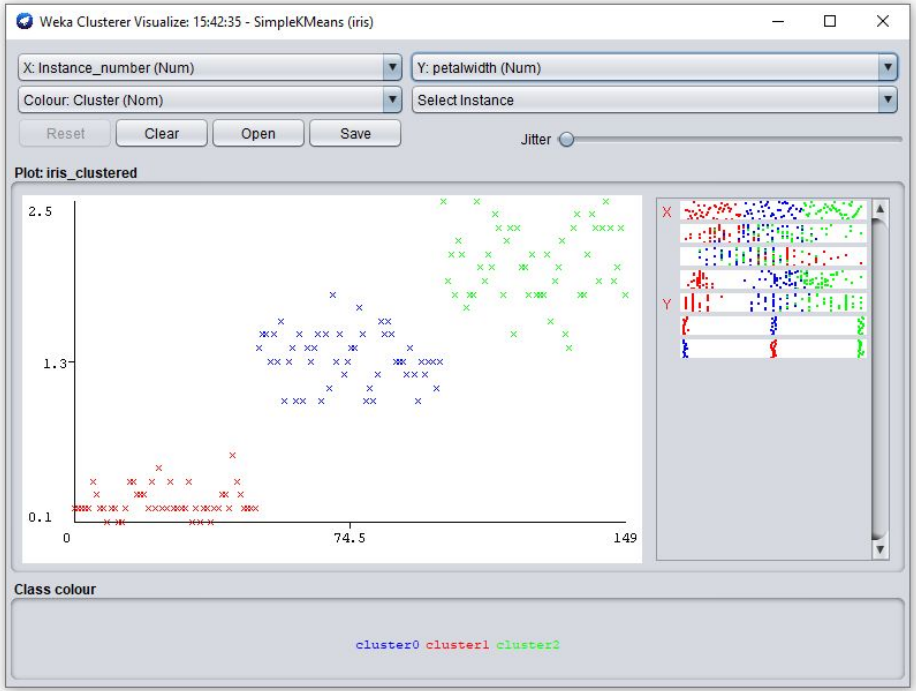
Scaled vs Non-scaled

```
pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_81dd79c73086
countByClusterLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cluster: int, class: string ... 1 more field]
+-------+--------------+-----+
|cluster|         class|count|
+-------+--------------+-----+
|      0|   Iris-setosa|   50|
|      1|Iris-versicolor|   50|
|      1| Iris-virginica|   50|
+-------+--------------+-----+

pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_558a2e3f6260
countByClusterLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cluster: int, class: string ... 1 more field]
+-------+--------------+-----+
|cluster|         class|count|
+-------+--------------+-----+
|      0|Iris-versicolor|   39|
|      0| Iris-virginica|   17|
|      1|Iris-versicolor|   11|
|      1| Iris-virginica|   33|
|      2|   Iris-setosa|   50|
+-------+--------------+-----+
```

Scaled vs Non-scaled

```
+-------+---------------+-----+
|cluster|          class|count|
+-------+---------------+-----+
|      0|Iris-versicolor|   47|
|      0| Iris-virginica|   50|
|      1|    Iris-setosa|   50|
|      1|Iris-versicolor|    3|
+-------+---------------+-----+

pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_6c50ed69fb29
countByClusterLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cluster: int, class: string ... 1 more field]
+-------+---------------+-----+
|cluster|          class|count|
+-------+---------------+-----+
|      0|    Iris-setosa|   50|
|      1|Iris-versicolor|   47|
|      1| Iris-virginica|   14|
|      2|Iris-versicolor|    3|
|      2| Iris-virginica|   36|
+-------+---------------+-----+
```

```
Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                    Cluster#
Attribute               Full Data           0               1               2
                          (150.0)        (50.0)          (50.0)          (50.0)
=============================================================================
sepallength                5.8433         5.936           5.006           6.588
sepalwidth                  3.054          2.77           3.418           2.974
petallength                3.7587          4.26           1.464           5.552
petalwidth                 1.1987         1.326           0.244           2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```



So why didn't this work out?

1. Data set was too sparse
2. The instances of versicolor and virginica was too similar
    a. Data may have been generated vs recorded

Tuning Iterations

cluster	class	count
0	Iris-setosa	50
1	Iris-versicolor	50
1	Iris-virginica	15
2	Iris-virginica	35

cluster	class	count
0	Iris-versicolor	46
0	Iris-virginica	13
1	Iris-setosa	50
2	Iris-versicolor	4
2	Iris-virginica	37

cluster	class	count
0	Iris-versicolor	48
0	Iris-virginica	14
1	Iris-setosa	50
2	Iris-versicolor	2
2	Iris-virginica	36

cluster	class	count
0	Iris-versicolor	50
0	Iris-virginica	15
1	Iris-setosa	50
2	Iris-virginica	35

cluster	class	count
0	Iris-versicolor	3
0	Iris-virginica	36
1	Iris-setosa	50
2	Iris-versicolor	47
2	Iris-virginica	14

cluster	class	count
0	Iris-versicolor	36
0	Iris-virginica	1
1	Iris-setosa	50
2	Iris-versicolor	14
2	Iris-virginica	49

cluster	class	count
0	Iris-versicolor	47
0	Iris-virginica	14
1	Iris-setosa	50
2	Iris-versicolor	3
2	Iris-virginica	36

cluster	class	count
0	Iris-setosa	50
1	Iris-versicolor	48
1	Iris-virginica	14
2	Iris-versicolor	2
2	Iris-virginica	36

cluster	class	count
0	Iris-versicolor	49
0	Iris-virginica	15
1	Iris-setosa	50
2	Iris-versicolor	1
2	Iris-virginica	35

cluster	class	count
0	Iris-setosa	50
1	Iris-versicolor	2
1	Iris-virginica	36
2	Iris-versicolor	48
2	Iris-virginica	14

```
(20,0.273021191057774,())
(40,0.273021191057774,())
(60,0.2895730091214502,())
(80,0.2895730091214502,())
(100,0.273021191057774,())
(120,0.2895730091214502,())
(140,0.2895730091214502,())
(160,0.2895730091214502,())
(180,0.273021191057774,())
(200,0.273021191057774,())
```

80 vs 1000 iterations

```
pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_7208d158d2ae
countByClusterLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cluster: int, class: string ... 1 more field]
+-------+---------------+-----+
|cluster|          class|count|
+-------+---------------+-----+
|      0|Iris-versicolor|    3|
|      0| Iris-virginica|   36|
|      1|    Iris-setosa|   50|
|      2|Iris-versicolor|   47|
|      2| Iris-virginica|   14|
+-------+---------------+-----+

pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_95a81387c5c6
countByClusterLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [cluster: int, class: string ... 1 more field]
+-------+---------------+-----+
|cluster|          class|count|
+-------+---------------+-----+
|      0|Iris-versicolor|   48|
|      0| Iris-virginica|   14|
|      1|    Iris-setosa|   50|
|      2|Iris-versicolor|    2|
|      2| Iris-virginica|   36|
+-------+---------------+-----+
```

The secret: setSeedRandom.nextlong

160 iterations

cluster	class	count
0	Iris-versicolor	6
0	Iris-virginica	44
1	Iris-setosa	50
2	Iris-versicolor	44
2	Iris-virginica	6

```
(20,0.273021191057774,())
(40,0.2895730091214502,())
(60,0.273021191057774,())
(80,0.273021191057774,())
(100,0.273021191057774,())
(120,0.273021191057774,())
(140,0.2895730091214502,())
(160,0.2895730091214502,())
(180,0.2895730091214502,())
(200,0.2895730091214502,())
```