Dokumentation K-Nearest-Neighbour Algorithmus

Wir haben uns entschlossen ein Framework zu bauen welches man für sämtliche Datensätze verwenden kann. Ich werde nun die Confusion-Matrix von dem Wein-Datensatz und für den IRIS-Datensatz. Zusätzlich haben wir entdeckt, dass der Algorithmus mit Java 8 ungefähr um das 10-fache schneller arbeitet als unter Java 9. Daher sind folgende Auswertungen mit Java 8 ausgewertet.

Man sieht, dass es beim Wein sehr schwierig ist zu Klassifizieren. Wenn man sich nun die Confusion-Matrix für die IRIS-Blüten (unten rechts) ansieht, erkennt man das die Accurancy auf 100 % ist. Für uns bedeutet es, dass die Daten vom Wein sehr schwierig auszuwerten sind.

```
Reading data...Done!
Removing outliers...Done!
Categorizing all datasets...Done!
Creating packs...Done!
Doing pass 1...Done!
Doing pass 2...Done!
Doing pass 3...Done!
Doing pass 4...Done!
Doing pass 5...Done!
Doing pass 6...Done!
Doing pass 7...Done!
Doing pass 8...Done!
Doing pass 9...Done!
Doing pass 10...Done!
Printing results...

6.0   5.0   7.0   8.0   4.0   3.0   9.0
-------------------------------------
1066   598   435   55   38   6   0
679   537   167   14   56   4   0
406   128   273   56   15   1   1
55   35   64   15   6   0   0
59   72   14   0   17   1   0
7   10   0   0   3   0   0
3   0   2   0   0   0   0
-------------------------------------
Vertical axis: Reference
Horizontal axis: Prediction
-------------------------------------
Accuracy: 38,95%
```

```
Reading data...Done!
Removing outliers...Done!
Categorizing all datasets...Done!
Creating packs...Done!
Doing pass 1...Done!
Doing pass 2...Done!
Doing pass 3...Done!
Doing pass 4...Done!
Doing pass 5...Done!
Doing pass 6...Done!
Doing pass 7...Done!
Doing pass 8...Done!
Doing pass 9...Done!
Doing pass 10...Done!
Printing results...

Iris-setosa   Iris-versicolor   Iris-virginica
-------------------------------------
50   0   0
0   50   0
0   0   50
-------------------------------------
Vertical axis: Reference
Horizontal axis: Prediction
-------------------------------------
Accuracy: 100%

Process finished with exit code 0
```

Nun haben wir mit dem Algorithmus die Zeit für 1000, 10 000, 100 000 Klassifizierungen gemessen und sind auf folgendes Ergebnis für die IRIS – Blüten gekommen:

```
Classified 1000 datasets in 0.383773387s!
```

```
Classified 10000 datasets in 2.936137456s!
```

```
Classified 100000 datasets in 26.687230296s!
```

Die Werte für den Wein-Datensatz für 1000, 10 000, 100 000 Datensätze ist folgender:

```
Classified 1000 datasets in 4.214601567s!
Classified 10000 datasets in 32.986494414s!
Classified 100000 datasets in 5m 18.096032038s!
```