

Clustering and Dimensionality Reduction

Biagio Alessandrello

2024-08-12

Clustering and dimensionality reduction

The data in `wine.csv` contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. In addition, two other variables about each wine are recorded: - whether the wine is red or white

- the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Run PCA, tSNE, and any clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes the most sense to you for this data? Convince yourself (and me) that your chosen approach is easily capable of distinguishing the reds from the whites, using only the “unsupervised” information contained in the data on chemical properties. Does your unsupervised technique also seem capable of distinguishing the higher from the lower quality wines? Present appropriate numerical and/or visual evidence to support your conclusions.

To clarify: I’m not asking you to run a supervised learning algorithms. Rather, I’m asking you to see whether the differences in the labels (red/white and quality score) emerge naturally from applying an unsupervised technique to the chemical properties. This should be straightforward to assess using plots.

PCA

- Standard deviation: This is the square root of the eigenvalue associated with PC1. It measures the
- Proportion of Variance: This indicates that PC1 explains % of the total variance in the data. This
- Cumulative Proportion: Since there’s only one principal component considered here, the cumulative p

```
## Loading required package: lattice
```

```
##
```

```
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
## lift
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
## 'data.frame': 6497 obs. of 13 variables:
```

```
## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
```

```
## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
```

```
## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
```

```
## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
```

```
## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
```

```
## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...
```

```
## $ total.sulfur.dioxide : num 34 67 54 60 34 40 59 21 18 102 ...
```

```
## $ density : num 0.998 0.997 0.997 0.998 0.998 ...
```

```
## $ pH          : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
## $ sulphates   : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
## $ alcohol     : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
## $ quality     : int   5 5 5 6 5 5 5 7 7 5 ...
## $ color       : chr   "red" "red" "red" "red" ...
```

```
## [1] 0
```

```
## Importance of components:
```

```
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion 0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##          PC8      PC9      PC10      PC11
## Standard deviation 0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion 0.94568 0.97632 0.9970 1.00000
```

PCA of Wine Data

