DSCI 100 - Introduction to Data Science

Lecture 12 - Visualizing high dimensional data & Data Science wrap-up

2019-04-04

Output of K means multivariate clustering from tutorial

```
K-means clustering with 3 clusters of sizes 123, 389, 288

Cluster means:
     Total       HP    Attack   Defense   Sp. Atk   Sp. Def    Speed
1 622.5691 88.91057 117.72358 100.65854 116.33333 101.86179 97.08130
2 472.9666 77.19280  85.30077  80.95373  77.54499  79.02057 72.95373
3 303.8958 50.14931  53.95486  52.78472  47.85417  49.49306 49.65972

Clustering vector:
  [1] 3 2 2 1 3 2 2 1 1 3 2 2 1 3 3 2 3 3 2 2 3 3 2 1 3 2 3 2 3 2 3 2 3 2 3 3 2
 [38] 3 3 2 3 2 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 3 2 3 2 3 2 3 1 3 3 2 3 2 2 1 3 2
 [75] 2 3 2 2 3 2 3 2 2 2 2 3 2 1 3 2 3 3 2 3 2 3 2 3 2 3 2 2 1 3 3 2 3 2 3 2 3
[112] 2 3 2 2 2 3 3 2 3 2 2 2 2 1 3 2 3 2 3 2 2 2 2 2 2 2 1 2 3 2 1 2 3 3 2 2 2
[149] 2 3 2 3 2 2 1 2 1 1 1 3 2 1 1 1 1 1 3 2 2 3 2 2 3 2 2 3 2 3 2 3 2 3 2 2 3
[186] 2 3 3 3 3 2 3 2 3 3 2 1 2 3 2 2 2 3 3 2 3 3 2 2 3 2 2 2 2 2 2 3 2 2 3 2 2
[223] 2 2 1 3 2 2 2 1 2 2 1 2 3 2 3 2 3 2 3 3 2 3 2 2 3 2 1 2 3 2 2 2 3 3 2 3 3
[260] 3 2 2 1 1 1 3 2 1 1 1 1 1 3 2 2 1 3 2 2 1 3 2 2 1 3 2 3 2 3 3 2 3 3 3 3 2
[297] 3 3 2 3 2 3 2 3 3 2 1 3 2 3 2 3 2 1 3 2 3 3 3 2 3 2 3 3 3 3 3 2 3 2 3 2 2
[334] 1 3 2 2 3 2 1 2 2 2 2 2 3 2 3 2 1 2 2 3 2 1 2 3 2 3 3 3 2 3 2 3 2 1 2 2 2
[371] 2 3 2 3 2 3 2 3 2 3 2 3 2 2 2 3 2 1 3 2 2 2 2 1 3 3 2 1 3 2 2 3 2 2 2 3 3
[408] 2 1 1 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 2 2 3 2 2 3 2 2 3 3 2
```

Output of K means multivariate clustering from tutorial

- we can look at total within sum of squares (but really only useful for comparing models)

- we can look at the ratio of between sum of squares / total sum of squares
  - if very small, then there are no discernable clusters
  - if 100, then each point is its own cluster

neither of these are very intuitive (at least to me)

What is intuitive?

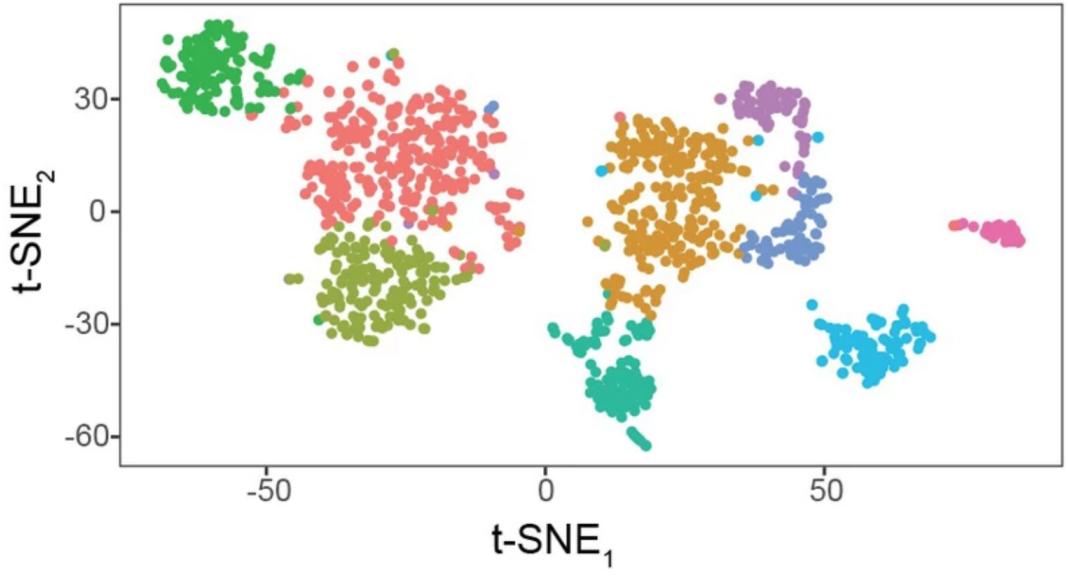
Visualization! A picture says 1000 words!

t-sne

- a popular dimensonality reduction algorithm useful for visualizing multi-dimensional data sets
- no "model" given from t-sne (only works to visualize the data you currently have)
- *see links in worksheet for more details about the specifics of the algorithm if you are interested*

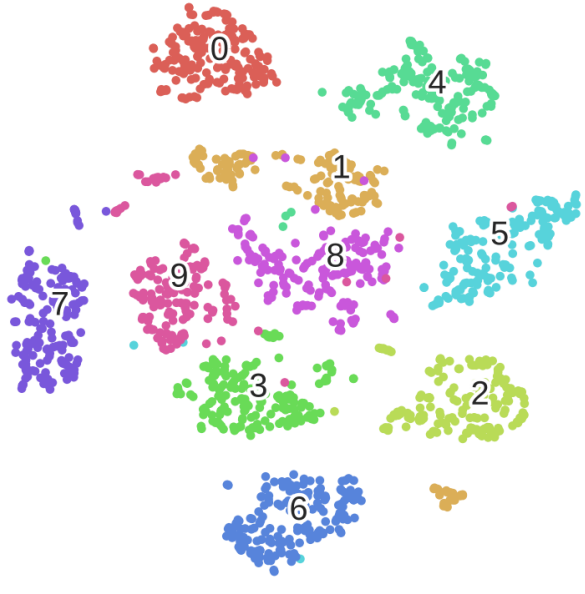
t-sne visualization of gene expression data from cells in a region of the brain

- each data point in this picture corresponds to a single brain cell for which we have the expression level measurements for thousands of genes.



t-sne visualization of hand-written digits data set overlaid with class identification

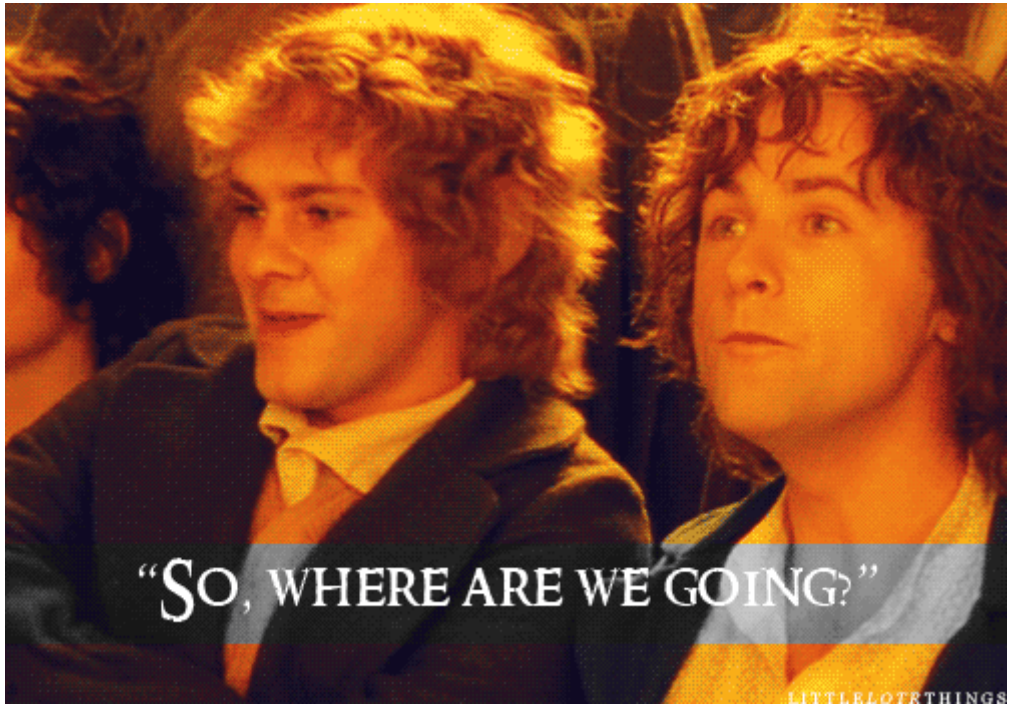
- each data point is an image of a handwritten digit for which we have 784 pixel values



COURSE EVALUATIONS!

Data Science wrap-up

In January, we started with this Gif



And we laid out these goals and this path:

High-level goals of this course:

1. Learn how to use reproducible tools (Jupyter + R) to do data analysis

1. Learn how to solve 3 common problems in Data Science

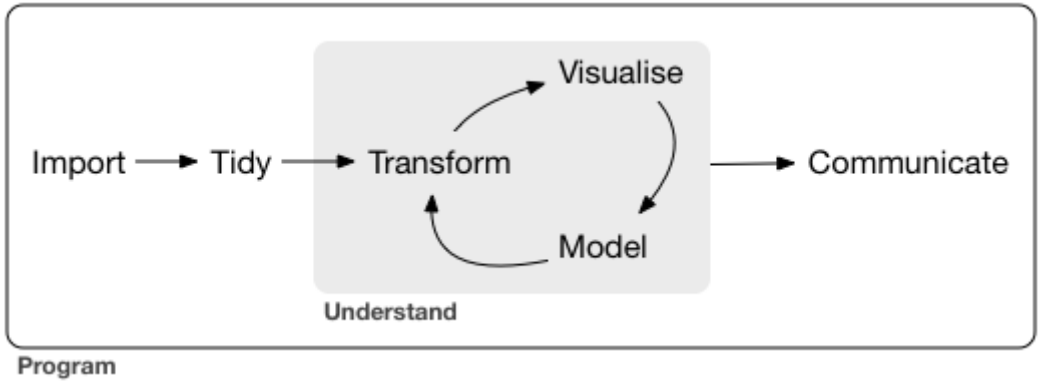
Problems we will focus on:

1. Predict a class/category for a new observation/measurement (*e.g.*, cancerous or benign tumour)

1. Predict a value for a new observation/measurement (*e.g.*, 10 km race time for a 35 year old with a BMI of 25).

1. Find previously unknown/unlabelled subgroups in your data (*e.g.*, products commonly bought together on Amazon)

Another way to think of what we did in this course:



source: R for Data Science (https://r4ds.had.co.nz/) by Grolemund & Wickham

Where to from here

- you learned a lot in this course!
- many of you are asking for more Data Science (yeah!)

- so here's a list of some UBC courses of interest you might want to take:
  - STAT 306 - Finding Relationships in Data (https://harlanhappydog.github.io/STAT306/)
  - STAT 406 - Methods for Statistical Learning (https://github.com/msalibian/STAT406)
  - CPSC 330 - Applied Machine Learning (Instructor coming to give a sneak peak today)
  - CPSC 340 - Machine Learning (https://www.cs.ubc.ca/~fwood/CS340/)
  - MATH 210 - Introduction to Mathematical Computing (https://github.com/ubc-math210/2018)

- outside of classes, I can recommend reading An Introduction to Statistical Learning (https://www-bcf.usc.edu/~gareth/ISL/) and the John Hopkins Coursera Data Science courses (https://www.coursera.org/specializations/jhu-data-science)

Thank-you and it's been a blast!