Machine learning, Statistics and Big Data

3-day short course

Participants: Australian Bureau of Statistics

Presenters: Hugh Anderson, Jacinta Holloway, Miles McBain, James McGree, Chris McCool, Kerrie Mengersen
Queensland University of Technology

October 2017

Program: Day 1

Time	Presentation
9.30am – 10.00am	Registration and Coffee
10.00am – 12.00pm	Welcome and overview of course 1. Overview of big data 2. Overview of stats & ML for big data: concepts, philosophy, terminology 3. Overview of computational frameworks: from divide & recombine to cloud computing 4. Case Study: grading images
12.00pm – 12.45pm	Lunch
12.45am – 2.45pm	1. Preparing your data 2. Overview of methods 3. Overview of algorithms
2.45pm – 3.00pm	Break
3.00pm – 4.30pm	Digging Deeper: Classification and Regression. 1. Generalised linear regression 2. Spatial and time series models 3. Tree-based approaches: CART, RF, BRT, bagging boosting 4. Support vector machines
4.30pm – 5.00pm	Extended Topics: Classification and Regression Semi-parametric regression, KNNs, Ensembles, XGBoost Discussion of cloud computing Concluding remarks: Day 1

Program: Day 2

Time	Presentation
9.30am – 10am	Coffee
10am – 12.00pm	Brief recap and discussion Digging Deeper: Clustering and Dimension Reduction. 1. kmeans 2. Mixture models 3. Feature extraction 4. PCA, FA and extensions 5. Page Rank
12.00pm – 12.45pm	Lunch
12.45am – 2.45pm	Digging Deeper: Neural networks 1. Overview of NNs 2. Convolutional and recurrent NNs 3. Deep learning
2.45pm – 3.00pm	Break
3.00pm – 4.30pm	Extended topics: NNs NNs for time series and 2D images
4.30pm – 5.00pm	Extended topics: NNs Deep Learning Systems Concluding remarks: Day 2

Program: Day 3

Time	Presentation
9.30am – 10.00am	Registration and Coffee
10.00am – 12.00am	Brief recap and discussion Case Study: Recommender systems. 1. Overview of recommender systems 2. Implementation 3. Use cases
12.00am – 12.45pm	Lunch
12.45pm – 2.45pm	The ABS context: Special Session. 1. Presentations from invited speakers 2. Discussion
2.45pm – 3.00pm	Break
3.00pm – 4.30pm	Extended Topics: 1. Overview of semi-supervised learning and ensembles of weak learners 2. Case study: return to classifying images
4.30pm – 5.00pm	Final issues Where to from here Concluding remarks: Day 3 Close

Day 2
Session 1

Digging Deeper: Clustering and Dimension Reduction.

1. kmeans
2. Mixture models
3. Feature extraction
4. PCA, FA and extensions
5. Page Rank

K-means

Partition observations into a fixed number (k) of clusters; each observation belongs to a specific cluster, based on similar properties.

Three steps:

1. Start the algorithm by choosing the number of clusters (k) and setting k points in the sample space. These points are chosen arbitrarily or according to some rule, and will be the initial cluster centres (centroids).
2. Allocate each observation to the closest cluster centroid.
3. Recalculate the cluster centroid as the mean, or average, of the observations that have been allocated to it.

Repeat steps 2 and 3 until the allocations stabilize.

The aim is to locate means and allocate observations to minimise within-cluster variation.

K-means

Given a set of observations

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

where each observation is a *d*-dimensional real vector,
partition the *n* observations into *k* (≤ *n*) sets

$$\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$$

to minimize the within-cluster sum of squares (WCSS)
(sum of distance functions of each point in the cluster to the K center).

So the objective is to find:

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu_i}||^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

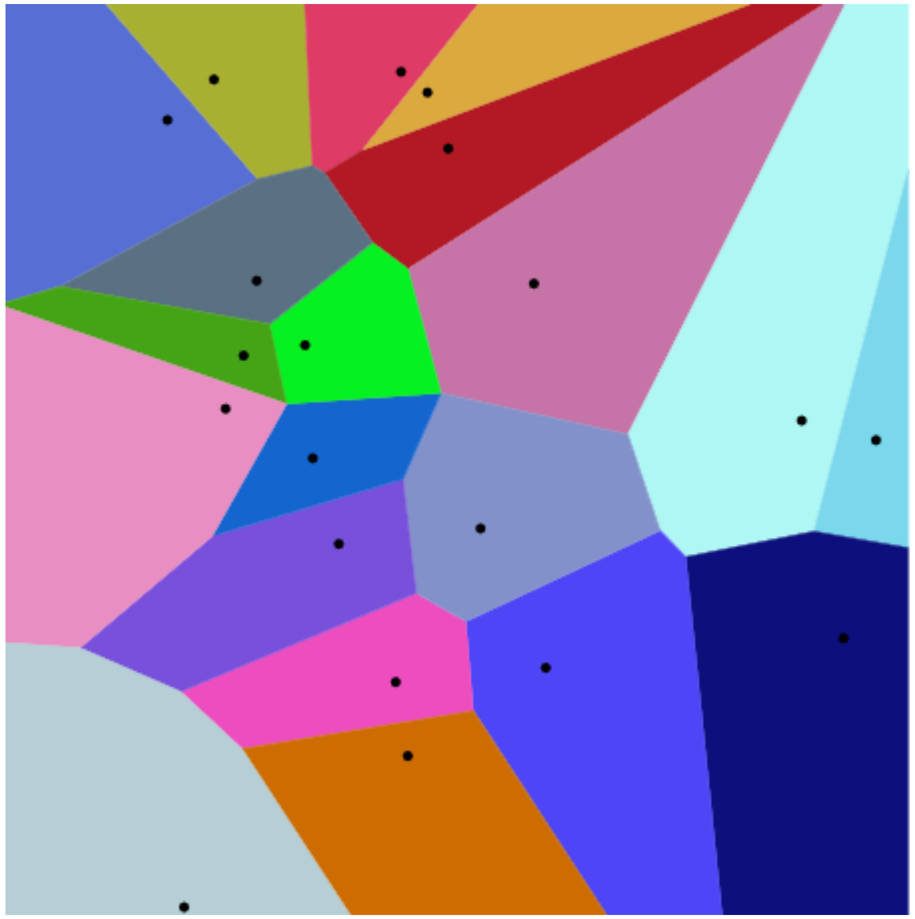
This also leads to a Voronoi partition.

Voronoi partitions

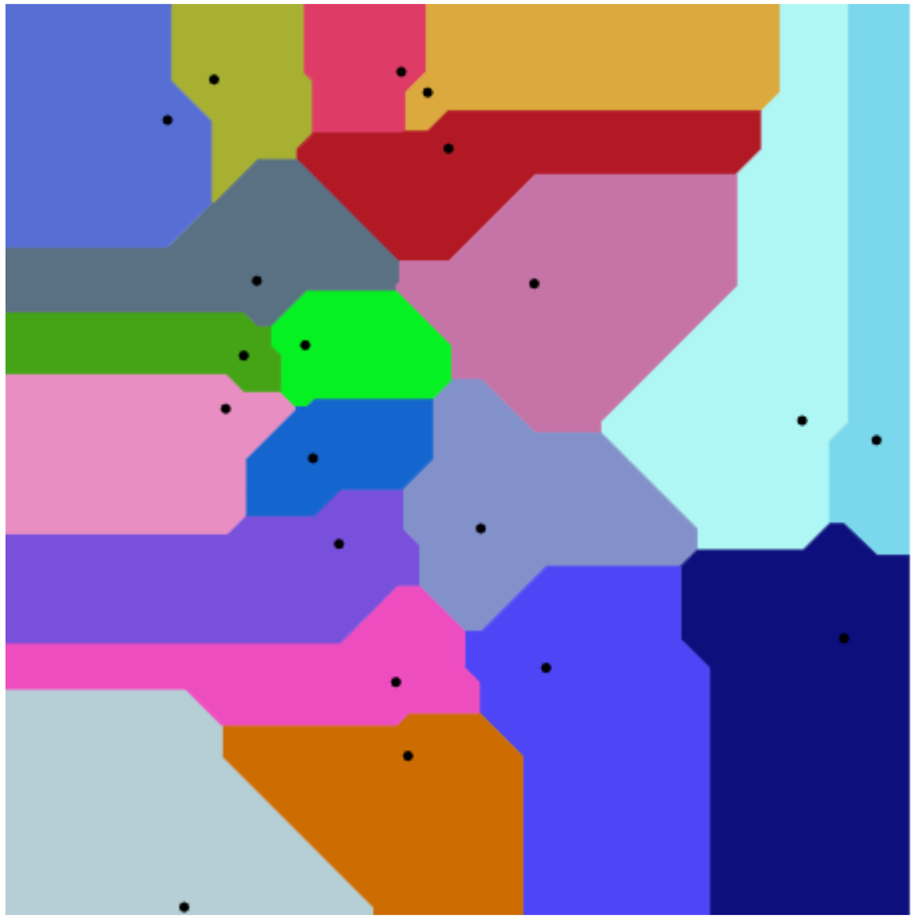
https://en.wikipedia.org/wiki/Voronoi_diagram

A **Voronoi diagram** is a partitioning of a plane into regions based on distance to points in a specific subset of the plane. That set of points (called seeds, sites, or generators) is specified beforehand, and for each seed there is a corresponding region consisting of all points closer to that seed than to any other. These regions are called Voronoi cells.

Eg: consider a group of shops in a city. Suppose we want to estimate the number of customers of a given shop. Assume that customers go to the nearest shop.
In this case the Voronoi cell of a given shop can be used for giving a rough estimate on the number of potential customers going to this shop
(which is modeled by a point in our city).
We can measure the distance between points using a range of distance metrics
Examples of metrics:
L2 (Euclidean) $d[(a_1, a_2), (b_1, b_2)] = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
L1 (Manhattan) $d[(a_1, a_2), (b_1, b_2)] = |a_1 - b_1| + |a_2 - b_2|$


Euclidean partition

Manhattan partition



https://en.wikipedia.org/wiki/Voronoi_diagram

Choosing k

- Sometimes there is a reason for specifying a certain number of clusters: this might be based on knowledge of the underlying biological or physical system in a scientific application, an economic or marketing rationale in business, and so on.
- If there is no such reason, then the k-means model can be run with different values of k, and the results compared.

Comparing results

- As k increases, the average distance between the observations and their cluster centroids decreases, which means that the observations within a cluster will be more similar.

- Sometimes there is a clear point at which the increase in k results in much less improvement in within-cluster similarity. The analyst can choose k to be at this point.

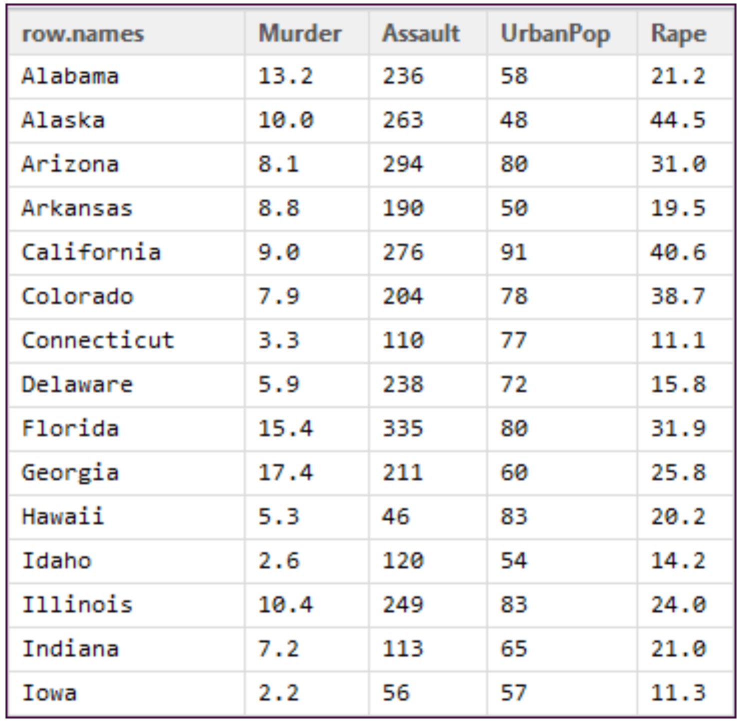
- Cross-validation can be used to provide a robust estimate of k. Cross-validation entails fitting the k-means algorithm to a subset of the data (called the training set) and then applying the clusters to the remaining data (the test set). This is particularly useful if the eventual aim is to allocate new observations to the clusters.

- The analyst can also gain insight into the effect of changing k by monitoring how particular observations are allocated to different groups.

Example

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

Crime data in 50 US states, per 100,000 people in a year



row.names	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

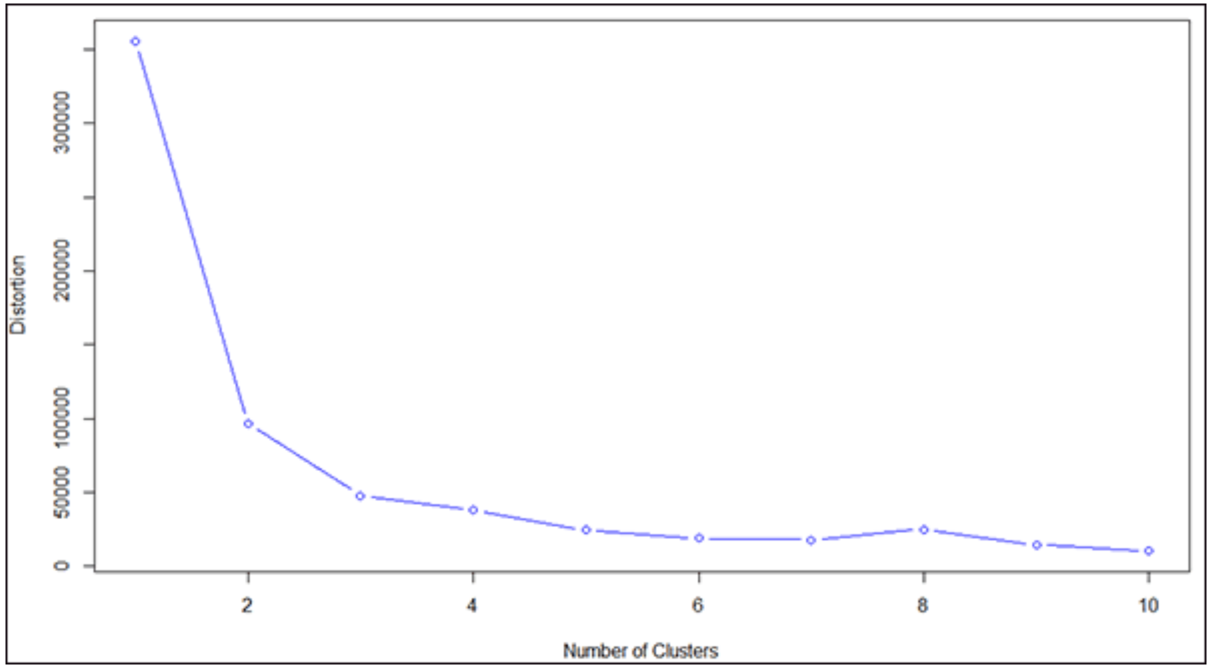
Results:

	crime$cluster	Murder	Assault	UrbanPop	Rape
Alabama	4	13.2	236	58	21.2
Alaska	4	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	3	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	3	7.9	204	78	38.7
Connecticut	2	3.3	110	77	11.1
Delaware	4	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9

```
No. clusters: 5
Total SS: 355808
Within SS: 4548, 2286, 1480, 3653
Total Within SS: 28240
Between SS: 327568
Size of clusters: 10, 9, 14, 10, 7
```

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

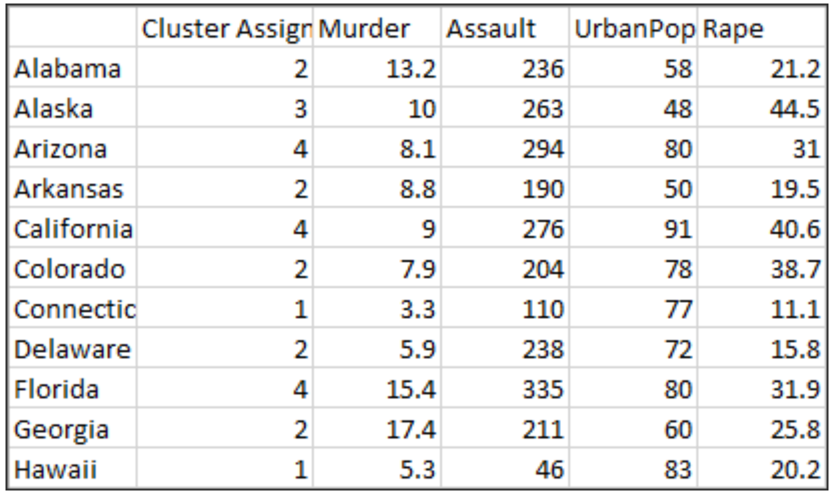
Compare distortion (via within SS) for various values of k:



https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

k=4:

```
Cluster centres:            Murder        Assault      UrbanPop  Rape
      Texas         4.74   104.85        62.96     16.10
      Louisiana    10.90   219.92        71.71     25.95
      South Carolina       13.37  284.50        46.25     25.05
      New Mexico   11.04   298.00        77.60     32.68
```

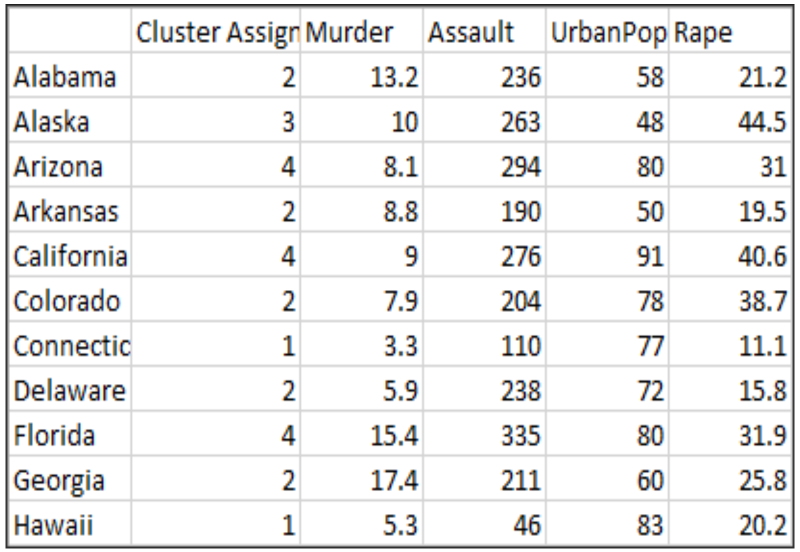


	Cluster Assign	Murder	Assault	UrbanPop	Rape
Alabama	2	13.2	236	58	21.2
Alaska	3	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	2	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	2	7.9	204	78	38.7
Connectic	1	3.3	110	77	11.1
Delaware	2	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9
Georgia	2	17.4	211	60	25.8
Hawaii	1	5.3	46	83	20.2

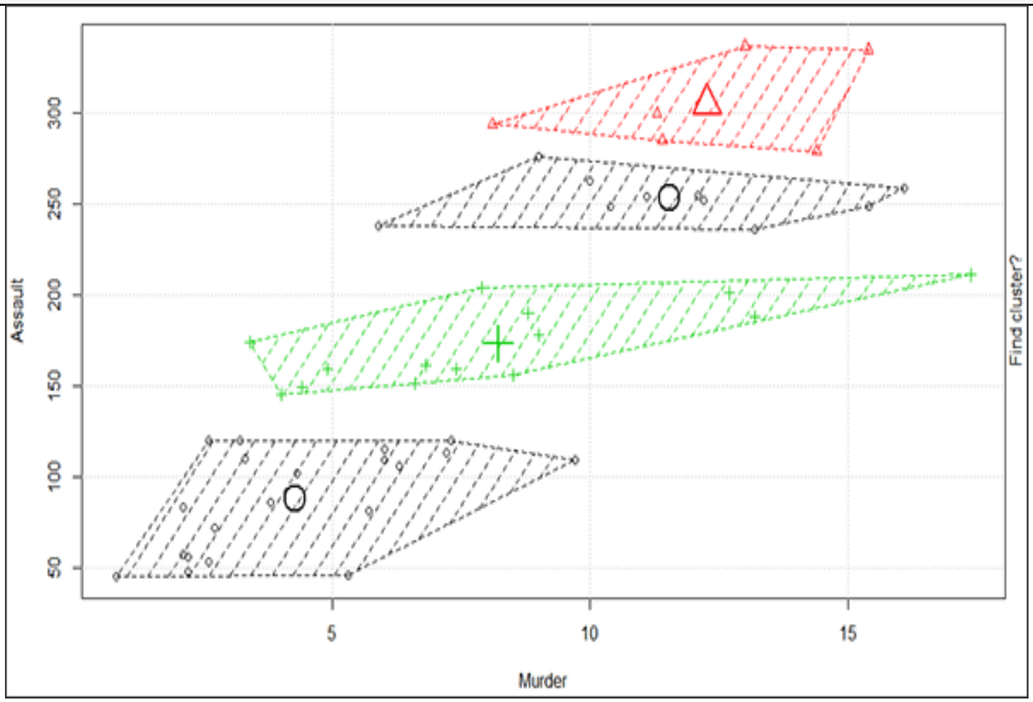
https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

k=4:

```
Cluster centres:      Murder          Assault       UrbanPop   Rape
 Texas          4.74   104.85          62.96       16.10
 Louisiana      10.90  219.92          71.71       25.95
 South Carolina        13.37  284.50          46.25       25.05
 New Mexico     11.04  298.00          77.60       32.68
```



	Cluster Assign	Murder	Assault	UrbanPop	Rape
Alabama	2	13.2	236	58	21.2
Alaska	3	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	2	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	2	7.9	204	78	38.7
Connectic	1	3.3	110	77	11.1
Delaware	2	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9
Georgia	2	17.4	211	60	25.8
Hawaii	1	5.3	46	83	20.2

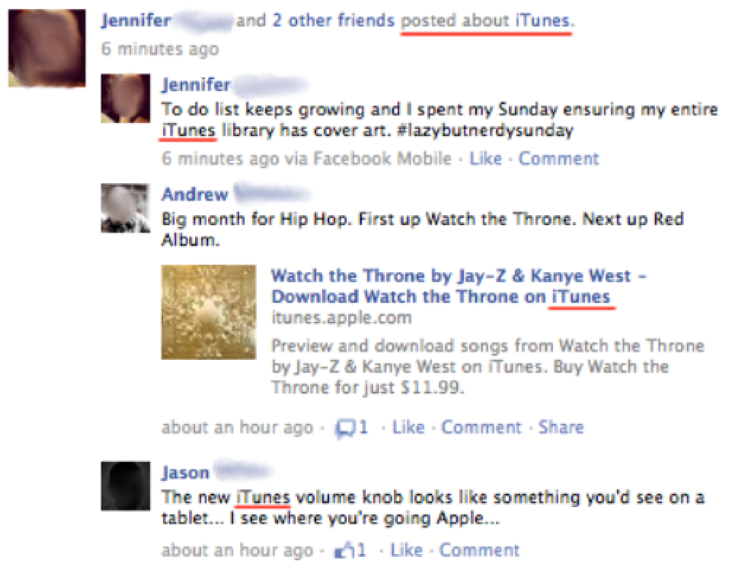


Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Users on Facebook

- ~300,000 status updates per minute on tens of thousands of topics
- Cluster users based on topic of status messages



Jennifer [redacted] and 2 other friends posted about iTunes.
6 minutes ago

Jennifer [redacted]
To do list keeps growing and I spent my Sunday ensuring my entire iTunes library has cover art. #lazybutnerdysunday
6 minutes ago via Facebook Mobile · Like · Comment

Andrew [redacted]
Big month for Hip Hop. First up Watch the Throne. Next up Red Album.

Watch the Throne by Jay-Z & Kanye West - Download Watch the Throne on iTunes
itunes.apple.com
Preview and download songs from Watch the Throne by Jay-Z & Kanye West on iTunes. Buy Watch the Throne for just $11.99.

about an hour ago · 1 · Like · Comment · Share

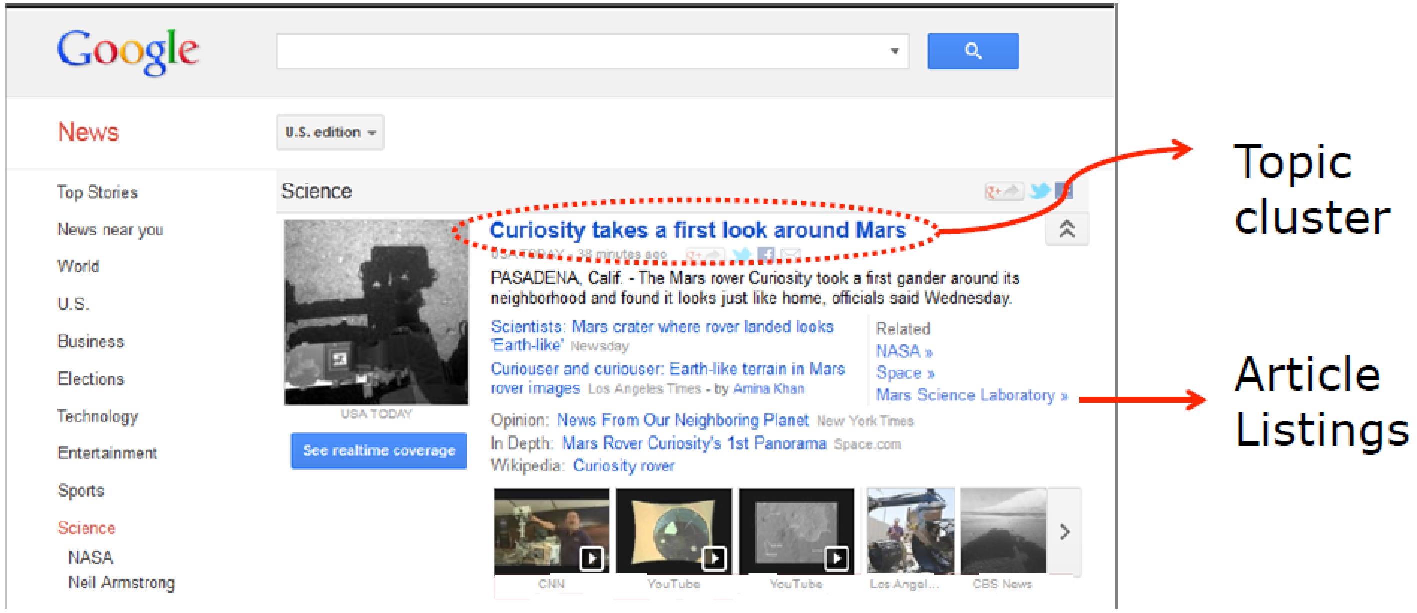
Jason [redacted]
The new iTunes volume knob looks like something you'd see on a tablet... I see where you're going Apple...
about an hour ago · 1 · Like · Comment

http://www.insidefacebook.com/2011/08/08/posted-about-page/
http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Articles on Google News



Curiosity takes a first look around Mars
USA TODAY - 38 minutes ago

PASADENA, Calif. - The Mars rover Curiosity took a first gander around its neighborhood and found it looks just like home, officials said Wednesday.

Scientists: Mars crater where rover landed looks 'Earth-like' Newsday

Curiouser and curiouser: Earth-like terrain in Mars rover images Los Angeles Times - by Amina Khan

Related
NASA »
Space »
Mars Science Laboratory »

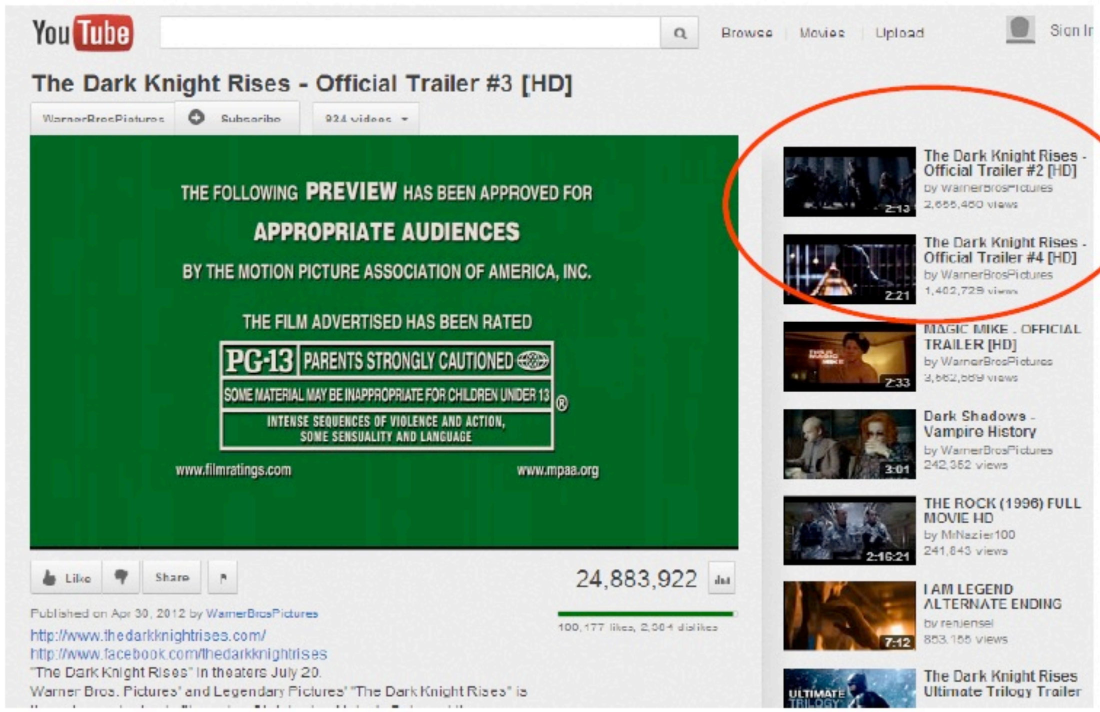
Opinion: News From Our Neighboring Planet New York Times
In Depth: Mars Rover Curiosity's 1st Panorama Space.com
Wikipedia: Curiosity rover

http://blogoscoped.com/archive/2006-07-28-n49.html

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Videos on Youtube



The Dark Knight Rises - Official Trailer #3 [HD]

THE FOLLOWING PREVIEW HAS BEEN APPROVED FOR APPROPRIATE AUDIENCES BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.

THE FILM ADVERTISED HAS BEEN RATED PG-13 PARENTS STRONGLY CAUTIONED SOME MATERIAL MAY BE INAPPROPRIATE FOR CHILDREN UNDER 13 INTENSE SEQUENCES OF VIOLENCE AND ACTION, SOME SENSUALITY AND LANGUAGE

www.filmratings.com www.mpaa.org

24,883,922

100,177 likes, 2,004 dislikes

Published on Apr 30, 2012 by WarnerBrosPictures
http://www.thedarkknightrises.com/
http://www.facebook.com/thedarkknightrises
"The Dark Knight Rises" in theaters July 20.
Warner Bros. Pictures' and Legendary Pictures' "The Dark Knight Rises" is

- 2:13 The Dark Knight Rises - Official Trailer #2 [HD] by WarnerBrosPictures 2,655,460 views
- 2:21 The Dark Knight Rises - Official Trailer #4 [HD] by WarnerBrosPictures 1,402,729 views
- 2:33 MAGIC MIKE - OFFICIAL TRAILER [HD] by WarnerBrosPictures 3,662,569 views
- 3:01 Dark Shadows - Vampire History by WarnerBrosPictures 242,352 views
- 2:16:21 THE ROCK (1996) FULL MOVIE HD by MrNazier100 241,843 views
- 7:12 I AM LEGEND ALTERNATE ENDING by renjensei 853,155 views
- The Dark Knight Rises Ultimate Trilogy Trailer

- Keywords
- Popularity
- Viewer engagement
- User browsing history

http://www.strutta.com/blog/blog/six-degrees-of-youtube

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering for Efficient Image retrieval

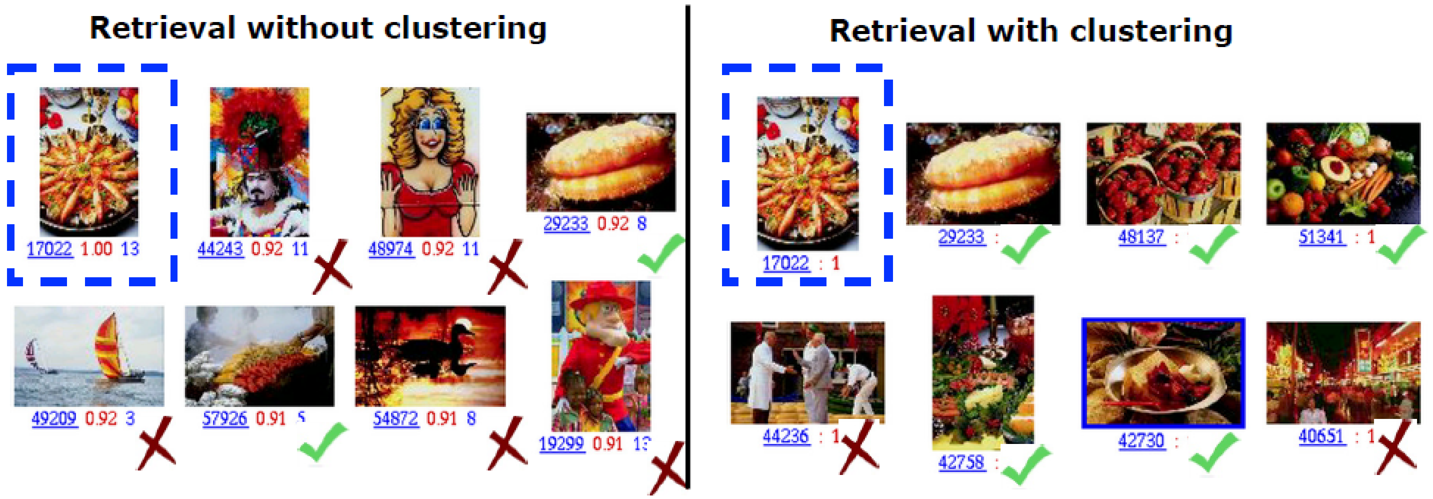


Fig. 1. Upper-left image is the query. Numbers under the images on left side: image ID and cluster ID; on the right side: Image ID, matching score, number of regions.

Retrieval accuracy for the "food" category (average precision):

Without clustering: **47%** With clustering: **61%**

Chen et al., "CLUE: cluster-based retrieval of images by unsupervised learning," IEEE Tans. On Image Processing, 2005.

When k-means clustering fails

https://dzone.com/articles/when-k-means-clustering-fails

When groups have very different sizes or data are categorical:

- Partitioning Around Medoids (**pam**)
- **clara** when the dataset is very large (and pam is slow)

While this [failure of k-means] may not be news to long-time clustering gurus it was a little sobering for us. It taught us once again that just because you've heard of some fancy algorithm and are using R's implementation of that algorithm does not guarantee that you'll get the results you expect. You always have to inspect results visually.

clara

https:

Consider sub-datasets of fixed size (sampsize) such that the time and storage requirements become linear in *n* rather than quadratic.

- Each sub-dataset is partitioned into k clusters using the same algorithm as in pam.
- Once k representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid.
- The mean (equivalent to the sum) of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the mean (or sum) is minimal, is retained.

A further analysis is carried out on the final partition.

- Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then.
- Randomly drawn observations are added to this set until sampsize has been reached.

https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clara.html

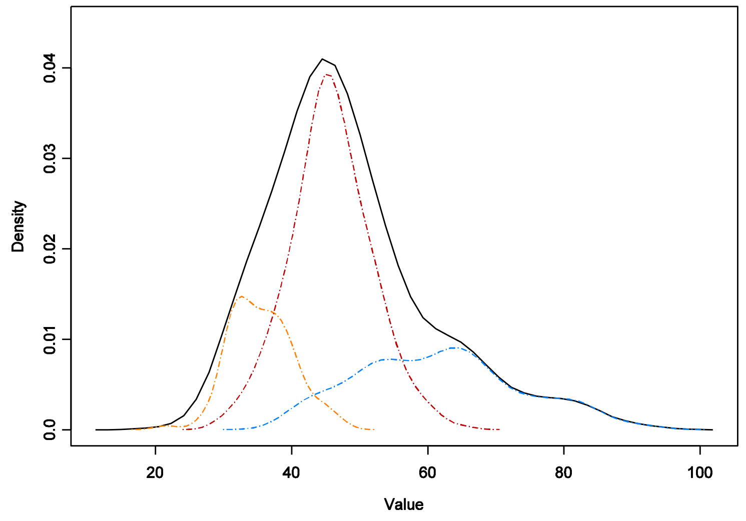
Day 2
Session 1

Digging Deeper: Clustering and Dimension Reduction.

1. kmeans
2. Mixture models
3. Feature extraction
4. PCA, FA and extensions
5. Page Rank

Mixture Models

The observed values
are observations from
a mixture of distributions



Eg, phenotypes from
3 genotypes: qq, qQ, QQ

Eg, for mixture of K=3 Normals: $\theta = (\mu,\sigma)$

$y \sim p_1 \, N(\mu_1, \sigma_1^2) + p_2 \, N(\mu_2, \sigma_2^2) + p_3 \, N(\mu_3, \sigma_3^2)$

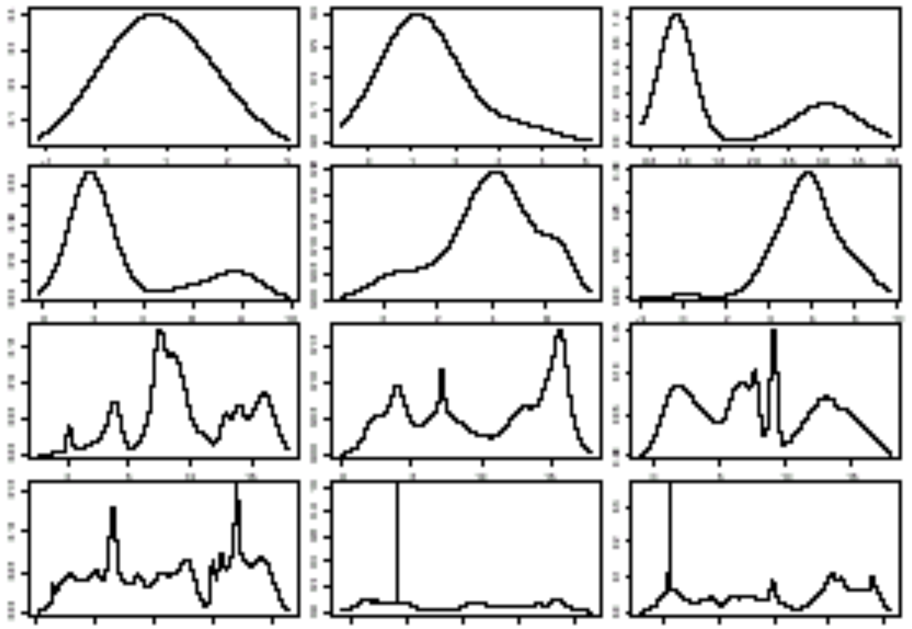


FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

Mixture models: Frequentist approach

Expectation-Maximisation (EM)

- expectation (E) step: computes the expectation of the log-likelihood evaluated using the current estimate for the parameters
- maximization (M) step: computes parameters maximizing the expected log-likelihood found on the *E* step.

For large datasets, use factor mixtures:

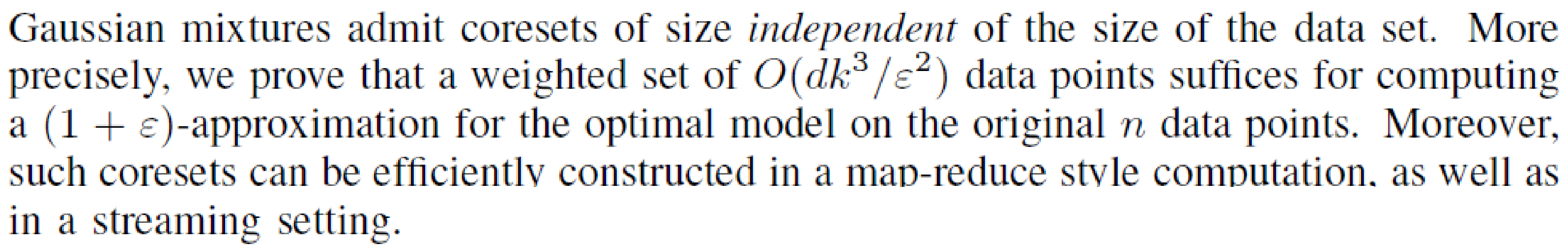
Lee, McLachlan, Pyne (2016), in *Big Data Analytics* by Pyne, Rao and Rao, Springer, 2016.

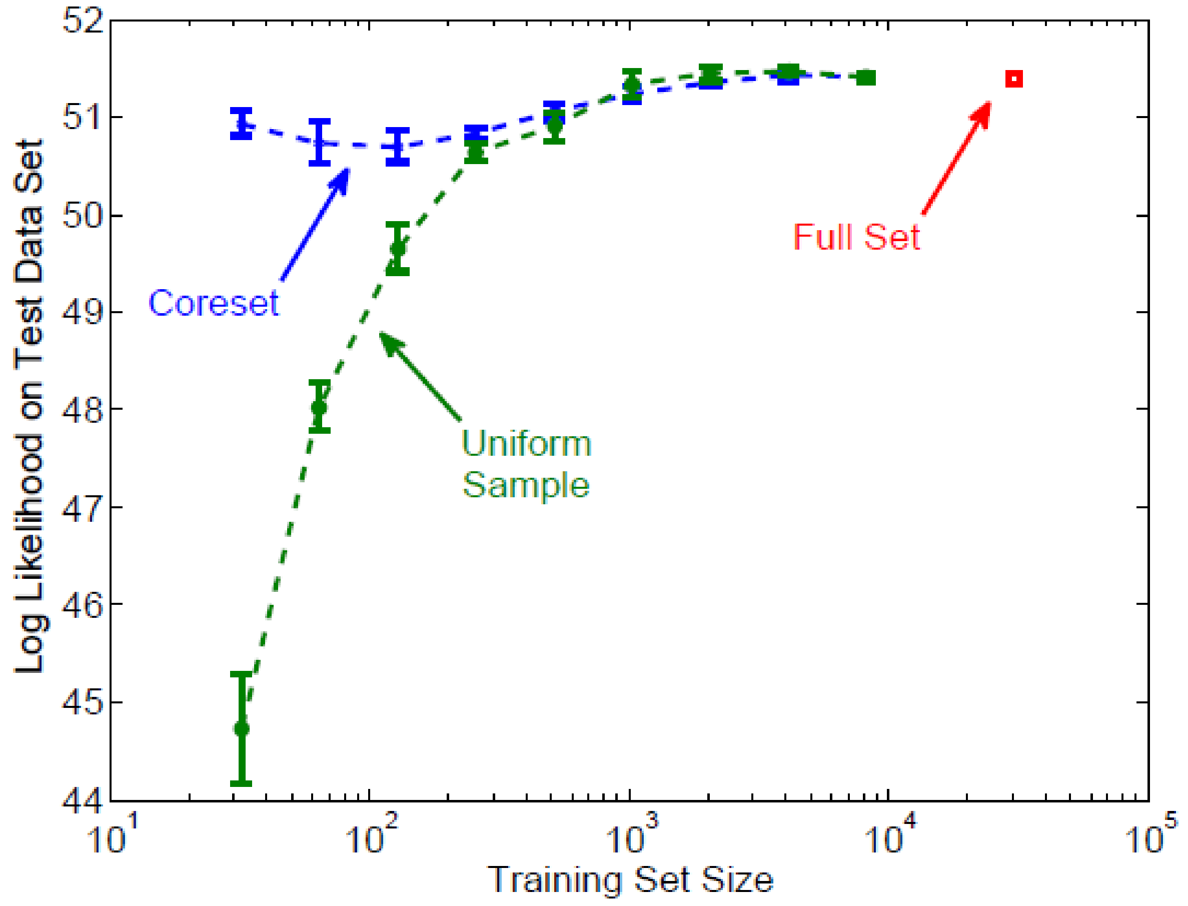
Mixture model approaches

Other two-step approaches

- Example: employ a two-step approach for very large datasets: (i) compress data by clustering the observations into a medium number of groups and representing each group by a triple of sufficient statistics (mean vector, covariance matrix, no. observations); (ii) estimate the mixture using by applying an adapted EM algorithm to the sufficient statistics; (iii) classify observations to clusters by maximum posterior probability of component membership (Steiner & Hudec, 2007).

- Example: use coresets: weighted subsets of the data (https://las.inf.ethz.ch/files/feldman11scalable-long.pdf)


Gaussian mixtures admit coresets of size *independent* of the size of the data set. More precisely, we prove that a weighted set of $O(dk^3/\varepsilon^2)$ data points suffices for computing a $(1+\varepsilon)$-approximation for the optimal model on the original n data points. Moreover, such coresets can be efficiently constructed in a map-reduce style computation, as well as in a streaming setting.


52
51
50
49
48
47
46
45
44
10^1
10^2
10^3
10^4
10^5
Log Likelihood on Test Data Set
Training Set Size
Coreset
Uniform Sample
Full Set

(a) *MNIST*

Bayesian approaches

$y \sim \Sigma_{j=1:k}\, p_j\, \mathrm{N}(\mu_j, \sigma_j^2)$

$\mu \sim$ Normal

$\sigma \sim$ Uniform

$p \sim$ Dirichlet$(\alpha_1,.., \alpha_K)$

$f(p;\alpha) \propto \Pi\, p_j^{\alpha_j-1}$; setting $\alpha=1$ for all j gives the Uniform.

Latent variable approach

- Associate with each y_i another variable T_i that identifies the component of the mixture to which that y_i belongs.
  (Note that we don't observe the *T*'s.)

- We can then 'break down' the likelihood:

$$y_i \mid T_i = T \sim \mathrm{N}(y \mid \mu_{Ti,} \sigma_{Ti}^2)$$

now just a univariate problem

- A typical prior for T is the multinomial or categorical distribution
  $T_i \sim Multi(p_1,\ldots,p_K)$

Gibbs sampling for mixtures

0. *Initialisation:* Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$ arbitrarily

```
    For t=1,...

    1.1 Allocate observations to components:
        Generate $T^{(t)}$ for each observation

    1.2 Generate new weights for the components:
        Generate $\underline{p}^{(t)}$

    1.3 Generate new parameters for each component:
        Generate $\underline{\theta}^{(t)}$
```

Day 2
Session 1

Digging Deeper: Clustering and Dimension Reduction.

1. kmeans
2. Mixture models
3. Feature extraction
4. PCA, FA and extensions
5. Page Rank

Feature extraction

- **Feature extraction** starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.
- Use for large and/or redundant input data: transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called *feature selection*.
- The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

https://en.wikipedia.org/wiki/Feature_extraction

Feature extraction techniques

- Independent component analysis
- Isomap
- Kernel PCA
- Latent semantic analysis
- Partial least squares
- Principal component analysis
- Multifactor dimensionality reduction
- Nonlinear dimensionality reduction
- Multilinear Principal Component Analysis
- Multilinear subspace learning
- Semidefinite embedding
- Autoencoder

https://en.wikipedia.org/wiki/Feature_extraction

Steps in feature extraction

Decompose the problem of feature extraction in two steps:

- feature construction
- feature selection

https://link.springer.com/content/pdf/10.1007/978-3-540-35488-8_1.pdf

Feature construction

- Finding a good set of features from “raw data” is domain-specific. Approaches range from human expertise to fully automatic methods.
- In some approaches, feature construction is integrated in the modeling process.
  - Example: the “hidden units” of artificial neural networks compute internal representations analogous to constructed features.
- In other approaches it is a pre-processing step. Preprocessing transformations may include:
  - Standardization: align different scales, eg by centring and scaling
  - Normalization: eg by computing proportions of pixels of interest by the total number of pixels in an image.
  - Signal enhancement. Improve the signal-to-noise ratio by applying signal or image-processing filters, eg baseline or background removal, de-noising, smoothing, or sharpening; popular examples are Fourier transform and wavelet transforms
  - Extraction of local features: eg using convolutional methods, often problem-specific
  - Linear and non-linear space embedding methods: useful when the dimensionality of the data is very high, to project or embed the data into a lower dimensional space while retaining as much information as possible, eg Principal Component Analysis (PCA) and Multidimensional Scaling (MDS)

https://link.springer.com/content/pdf/10.1007/978-3-540-35488-8_1.pdf

Feature selection

Feature selection is primarily performed to select relevant and informative features.

- It can have other motivations, including:
  - general data reduction, to limit storage requirements and increase algorithm speed
  - feature set reduction, to save resources in the next round of data collection or during utilization
  - performance improvement, to gain in predictive accuracy
  - data understanding, to gain knowledge about the process that generated the data or simply visualize the data

https://link.springer.com/content/pdf/10.1007/978-3-540-35488-8_1.pdf

Other things to know about feature extraction

- Filter methods: often identified to feature ranking methods.
  Such methods provide a complete order of the features using a relevance index.
  Methods for computing ranking indices include correlation coefficients, classical test statistics (t-test, F-test, chi-squared, etc.)
  More generally, methods that select features without optimizing the performance of a predictor are referred to as “filters”.

- Wrappers and embedded methods: these methods involve the predictor as part of the selection process.
  - Wrappers utilize a learning machine as a “black box” to score subsets of features according to their predictive power.
  - Embedded methods perform feature selection in the process of training and are usually specific to given learning machines.

https://link.springer.com/content/pdf/10.1007/978-3-540-35488-8_1.pdf

Others have different views of definitions…

Feature extraction: transform arbitrary data, such as text or images, into numerical features usable for machine learning.

Feature selection: a machine learning technique applied on these features.

http://scikit-learn.org/stable/modules/feature_extraction.html

Day 2
Session 1

Digging Deeper: Clustering and Dimension Reduction.

1. kmeans
2. Mixture models
3. Feature extraction
4. PCA, FA and extensions
5. Page Rank

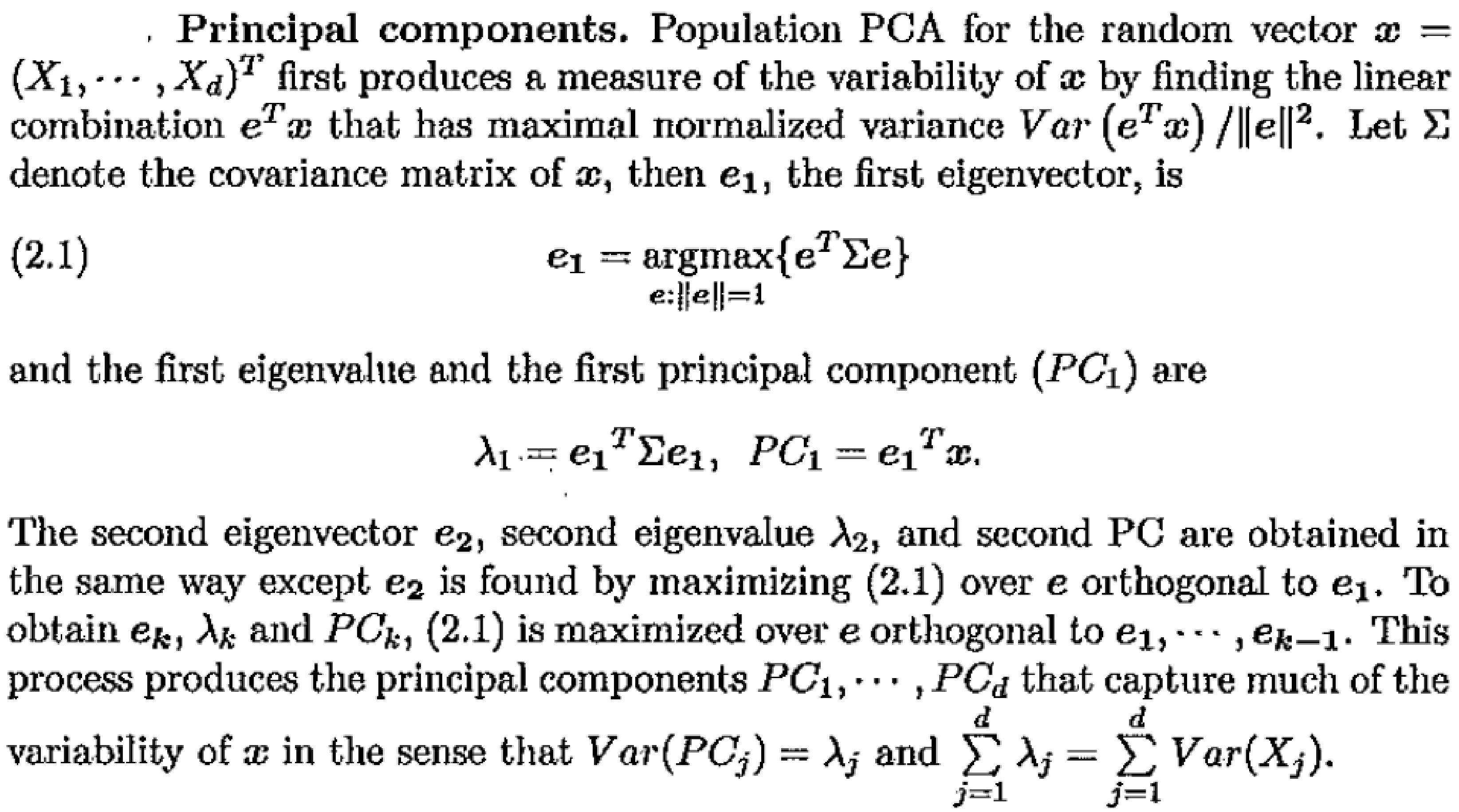
PCA

- Principal components analysis, or PCA, seeks to find a set of orthogonal axes such that the first axis, or first principal component, accounts for as much variability as possible and subsequent axes are chosen to maximize variance while maintaining orthogonality with previous axes.

- Principal components are typically computed either by a singular value decomposition of the data matrix or an eigenvalue decomposition of a covariance or correlation matrix.

https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

Eigenvalue analysis



Principal components. Population PCA for the random vector $x = (X_1, \cdots, X_d)^T$ first produces a measure of the variability of x by finding the linear combination $e^T x$ that has maximal normalized variance $Var\left(e^T x\right)/\|e\|^2$. Let Σ denote the covariance matrix of x, then e_1, the first eigenvector, is

$$e_1 = \underset{e:\|e\|=1}{\operatorname{argmax}}\{e^T \Sigma e\} \tag{2.1}$$

and the first eigenvalue and the first principal component (PC_1) are

$$\lambda_1 = {e_1}^T \Sigma e_1, \quad PC_1 = {e_1}^T x.$$

The second eigenvector e_2, second eigenvalue λ_2, and second PC are obtained in the same way except e_2 is found by maximizing (2.1) over e orthogonal to e_1. To obtain e_k, λ_k and PC_k, (2.1) is maximized over e orthogonal to $e_1, \cdots, e_{k-1}$. This process produces the principal components $PC_1, \cdots, PC_d$ that capture much of the variability of x in the sense that $Var(PC_j) = \lambda_j$ and $\sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{d} Var(X_j)$.

http://www.stat.wisc.edu/~doksum/papers/BigData.pdf

Example

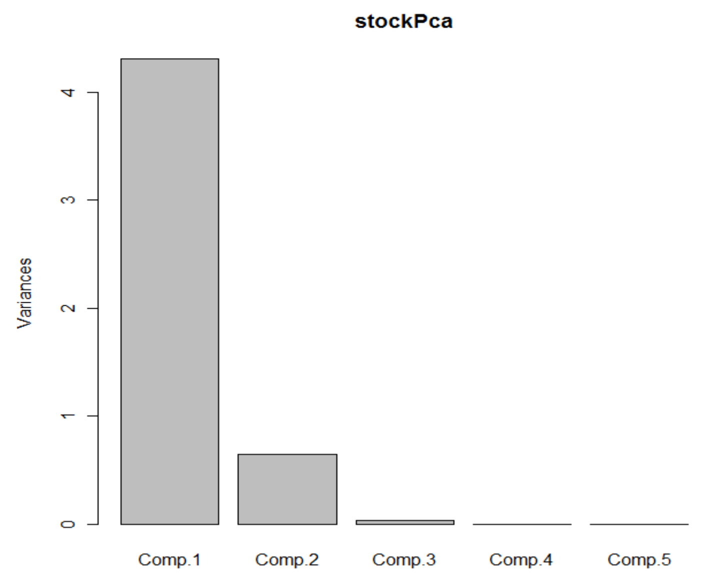
- Eg: Stock market data for open, high, low, close, and adjusted close from 1962 to 2010: 9.2 million observations of daily data for 2800 stocks.

```
➢ summary(stockPca)
➢ Importance of components: Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
➢ Standard deviation         2.0756 0.8063 0.1976 0.0454 0.0018
➢ Proportion of Variance     0.8616 0.1300 0.0078 0.0004 6.7E-5
➢ Cumulative Proportion      0.8616 0.9917 0.9995 0.9999 1.0000
```

```
➢ loadings(stockPca)
➢ Loadings:                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
➢ stock_price_open           -0.470 -0.166 0.867
➢ stock_price_high           -0.477 -0.151 -0.276 0.410 -0.711
➢ stock_price_low            -0.477 -0.153 -0.282 0.417 0.704
➢ stock_price_close          -0.477 -0.149 -0.305 -0.811
➢ stock_price_adj_close -0.309 0.951
```

https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

Example



https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

See also
http://www.bigdatanews.com/profiles/blogs/principal-component-analysis-using-r

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Day 2
Session 1

Digging Deeper: Clustering and Dimension Reduction.

1. kmeans
2. Mixture models
3. Feature extraction
4. PCA, FA and extensions
5. Page Rank

PageRank

- Aims to rank web pages based on their hyperlinks (i.e., links between the pages).
- This algorithm underpins the search engine Google, and variations of the algorithm are now used for every online search engine.
- A hyperlink from page x to page y is defined as a vote, by page x, for page y. Votes casted by pages that are themselves "important" weigh more heavily and help to make other pages more "important". This is exactly the idea of rank prestige in social networks.
- The computation of PageRank values of the Web pages can be done using the power iteration method, which produces a principal eigenvector with an eigenvalue of 1. The iteration ends when the PageRank values do not change much (e.g, the sum of the absolute values of the residuals are less than a specified threshold).

See also https://en.wikipedia.org/wiki/PageRank

http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html