+Babbel

User segmentation

Ken Schröder
data scientist (product)

21 FEB 2020

Agenda

User segmentation

Start the project
Dataset
Clustering

K-Means

Synthetic example
The algorithm
Challenges

Business value

Communication
Long run value

PCA

Curse of dimensionality
Correlation

User segmentation

Starting up the project

- Aim: Identify groups of users
- Non data-driven attempts
- Data driven approach

User segmentation

Getting the data

- User based
- Everything derived from event data
- 12 week interval
- Subscribed users only

User segmentation

Got the data, what’s next?

- Clustering!
- Which method to use?
- Dive into K-Means >_

Agenda

User segmentation

Start the project

Dataset

Clustering

K-Means

Synthetic example

The algorithm

Challenges

Business value

Communication

Long run value

PCA

Curse of dimensionality

Correlation

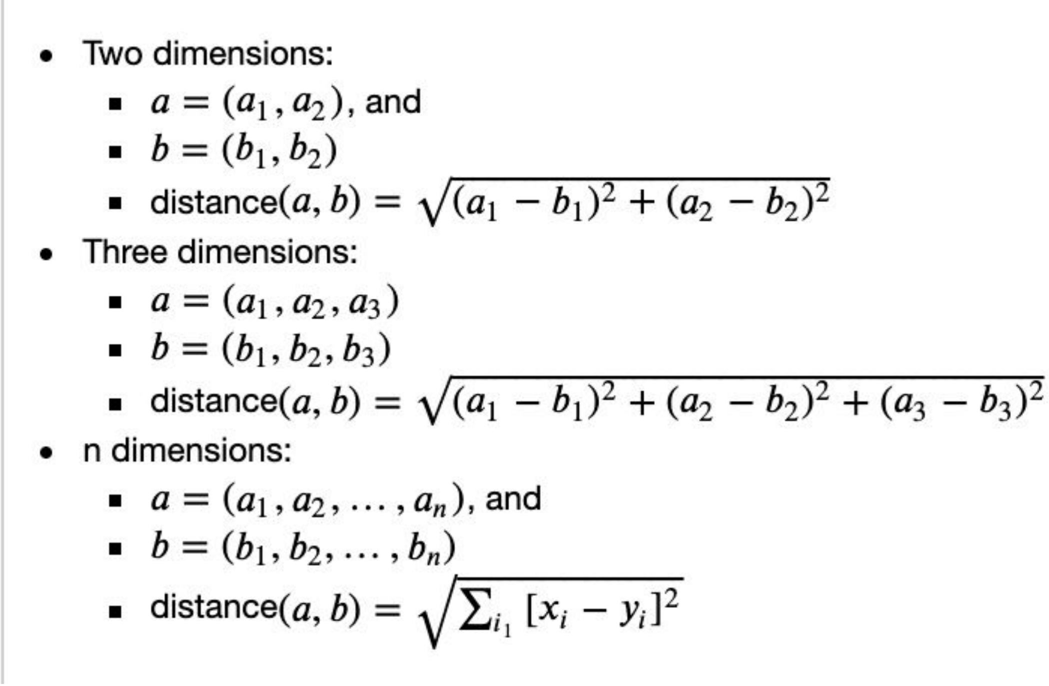
Clustering

Challenges

- More than 2 dimensions
- Variable 1 in (0, 1)   Variable 2 in (0, infinity)
- How many clusters to identify?
- Duplicate information in dataset

Clustering

More than 2 dimensions



- Two dimensions:
  - $a = (a_1, a_2)$, and
  - $b = (b_1, b_2)$
  - distance$(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
- Three dimensions:
  - $a = (a_1, a_2, a_3)$
  - $b = (b_1, b_2, b_3)$
  - distance$(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$
- n dimensions:
  - $a = (a_1, a_2, \ldots, a_n)$, and
  - $b = (b_1, b_2, \ldots, b_n)$
  - distance$(a, b) = \sqrt{\sum_{i_1} [x_i - y_i]^2}$

Clustering

Scaling

- What if we treat all distance the same?
- Need the same scale
- Common standardisation:
    - Deduct average
    - Divide by standard deviation

Clustering

Determining K

- Business aspect:
  - Interpretability: Centroids as real users
  - Simple

- Clustering quality metrics:
  - Similarity within cluster
  - Differences between cluster
  - E.g. silhouette score

Beware of effects of changing K, even by one

Agenda

User segmentation

Start the project

Dataset

Clustering

K-Means

Synthetic example

The algorithm

Challenges

Business value

Communication

Long run value

PCA

Curse of dimensionality

Correlation

Business value

Communication

- Centroid = average value of all its members
- I have 100+ variables!
- Find distinctive and interpretable dimensions
- Create a story, name clusters

Business value

Long run value (1/3)

- Should we retrain model?
- New clusters each iteration?
- User groups may become inconsistent

Business value

Long run value (2/3)

- Train K-Means once
- Pickle files
- Use pre-trained model for new data

Business value

Long run value (3/3)

- Job running that collects data weekly
- Extract and transform
- Save intermediate data in database
- Query into relevant data structure
- Apply pre-trained models & add to database tables

Agenda

User segmentation

Start the project

Dataset

Clustering

K-Means

Synthetic example

The algorithm

Challenges

Business value

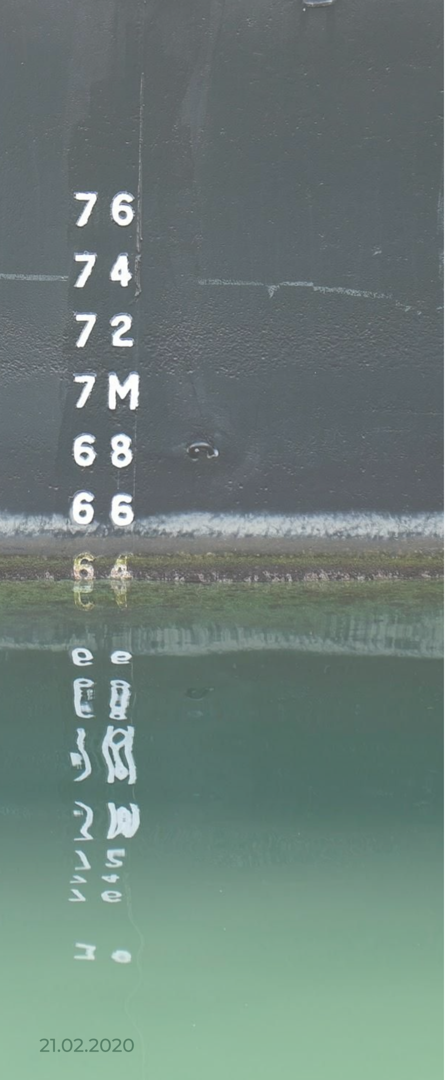
Communication

Long run value

PCA

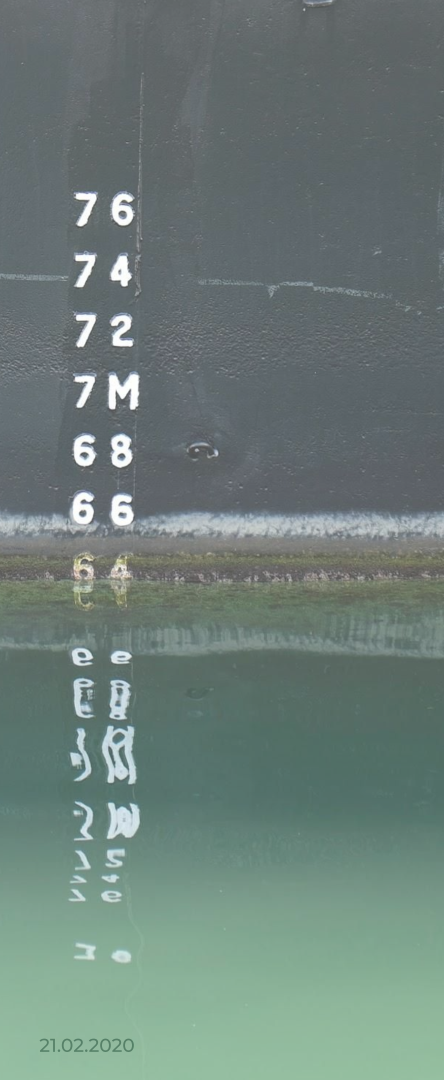
Curse of dimensionality

Correlation



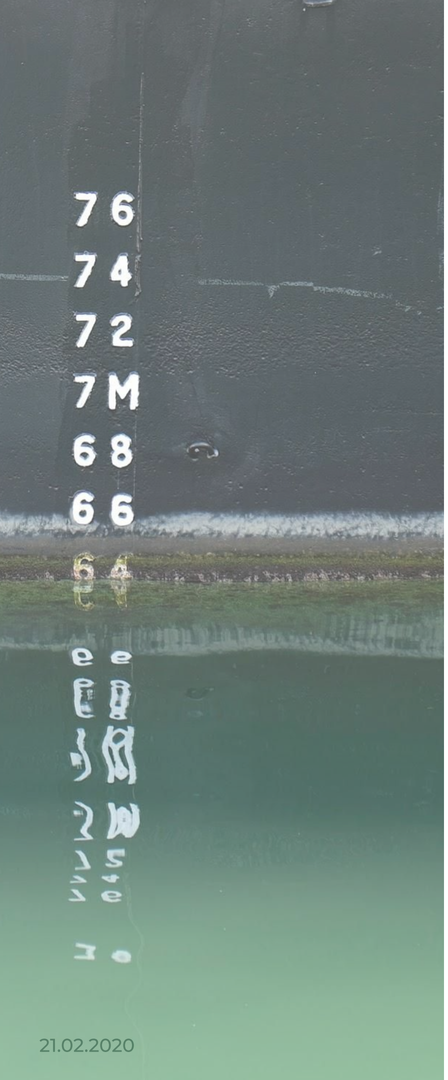
Principal component analysis

- High dimensionality
    - Distances become indistinguishable
    - Poor discrimination between the nearest and furthest neighbor
- Correlation
    - Same information, measured multiple times
    - Will dominate distance measure



Principal component analysis

- Use mathematical concepts like eigenvalues and eigenvectors
- Result is a rearrangement of the axes
- Linear transformation of the data
- Axes represent principal components



Principal component analysis

Challenges

- How many principal components?
- How to interpret resulting clusters?

Thank you!

Babbel