Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 20: Distance models (clustering)

Apr 6, 2017
Clustering

• Clustering (and unsupervised learning more generally) finds structure in data, using just $X$

$X = \text{a set of skyscrapers}$
Flat Clustering

- Partitions the data into a set of $K$ clusters
K-means

Algorithm 1 K-means

1: Data: training data \( x \in \mathbb{R}^F \)
2: Given some distance function \( d(x, x') \to \mathbb{R} \)
3: Select \( k \) initial centers \( \{\mu_1, \ldots, \mu_k\} \)
4: while not converged do
5: \hspace{1em} for \( i = 1 \) to \( N \) do
6: \hspace{2em} Assign \( x_i \) to \( \arg \min_c d(x_i, \mu_c) \)
7: \hspace{1em} end for
8: \hspace{1em} for \( i = 1 \) to \( K \) do
9: \hspace{2em} \( \mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i \)
10: \hspace{1em} end for
11: end while
Visualizing K-Means Clustering

http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html
Problems
K-means

initial cluster centers
K-means++

• Improved initialization method for K-means:
  • Choose data point at random as first center
  • For all other data points $x$, calculate the distance $D(x)$ between $x$ and the nearest cluster center
  • Choose new data point $x$ as next center, with probability proportional to $D(x)^2$
  • Repeat until $K$ centers are selected
K-means++

D(x)^2 = 1

1

10

D(x)^2 = 100

D(x)^2 = 1

D(x)^2 = 121
Choosing $K$

- how do we choose $K$?
The “elbow”

Core idea: clusters should minimize the within-cluster variance
The “elbow”

Core idea: clusters should minimize the within-cluster variance

within-cluster sum of squares

$$\sum_{i=1}^{F} (x_i - \mu_i)^2$$

for each cluster
The “elbow”
Gap statistic

• How much variance should we expect to see for a given number of clusters?

• Choose number of clusters that maximizes the “gap” between the observed variance and the expected variance for a given $K$.

Tibshirani et al., “Estimating the number of clusters in a data set via the gap statistic”
Hierarchical clustering

Core idea: build a binary tree of a set of data points by repeatedly merging the two most similar elements.
Hierarchical clustering

Algorithm 1 Hierarchical agglomerative clustering

1: Data: $N$ training data points $x \in \mathbb{R}^F$
2: Let $X$ denote a set of objects $x$
3: Given some linkage function $d(X, X') \to \mathbb{R}$
4: Initialize clusters $\mathcal{C} = \{C_1, \ldots, C_N\}$ to singleton data points
5: while data points not in one cluster do
6: Identify $X, Y$ as clusters with smallest linkage function among clusters in $\mathcal{C}$
7: Create new cluster $Z = X \cup Y$
8: remove $X, Y$ from $\mathcal{C}$
9: add $Z$ to $\mathcal{C}$
10: end while
Hierarchical clustering

Allison et al. 2009
Hierarchical clustering

We know how to compare data points with distance metrics.

How do we compare sets of data points?
Single linkage

\[
\min_{x \in A, \ y \in B} \text{Dis}(x, y)
\]
Complete linkage

$\max_{x \in A, y \in B} \text{Dis}(x, y)$
Average linkage

\[ \frac{\sum_{x \in A, y \in B} \text{Dis}(x, y)}{|A| \times |B|} \]
Single linkage may link bigger clusters together before outliers
Complete linkage may *not* link close clusters together because of outliers.
• Allison et al., “Quantitative Formalism: an Experiment”
### Dictionary mapping ngrams to classes

<table>
<thead>
<tr>
<th>First Person</th>
<th>Numbers</th>
<th>Positivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>about me</td>
<td>six-wheeled</td>
<td>perpetual adorations</td>
</tr>
<tr>
<td>about my</td>
<td>275 degrees</td>
<td>mated with</td>
</tr>
<tr>
<td>am</td>
<td>three-card loo</td>
<td>hugging yourself</td>
</tr>
<tr>
<td>I</td>
<td>695</td>
<td>striking responsive cord</td>
</tr>
<tr>
<td>I'd</td>
<td>four-ply</td>
<td>wassailing</td>
</tr>
<tr>
<td>I'll</td>
<td>half-way</td>
<td>plucked up your spirits</td>
</tr>
<tr>
<td>I'm</td>
<td>three parts</td>
<td>offers ourselves</td>
</tr>
<tr>
<td>I for one</td>
<td>eight-member</td>
<td>promotive of</td>
</tr>
<tr>
<td>ich</td>
<td>third-world</td>
<td>enshrining</td>
</tr>
<tr>
<td>ich dien</td>
<td>3,5</td>
<td>devotes yourself</td>
</tr>
<tr>
<td>me</td>
<td>half-and-half measures</td>
<td>music lover</td>
</tr>
<tr>
<td>mea</td>
<td>8,3</td>
<td>delectated</td>
</tr>
<tr>
<td>meum</td>
<td>half-reclining</td>
<td>recharging my batteries</td>
</tr>
<tr>
<td>mine</td>
<td>26</td>
<td>recommends you for</td>
</tr>
<tr>
<td>my</td>
<td>634</td>
<td>shadow of your smile</td>
</tr>
<tr>
<td>myself</td>
<td>five-rater</td>
<td>regaining our composure</td>
</tr>
</tbody>
</table>
Only unigrams with relative frequency > 0.03

<table>
<thead>
<tr>
<th>a</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>of</td>
</tr>
<tr>
<td>and</td>
<td>on</td>
</tr>
<tr>
<td>as</td>
<td>p_ahos</td>
</tr>
<tr>
<td>at</td>
<td>p_comma</td>
</tr>
<tr>
<td>be</td>
<td>p_exlam</td>
</tr>
<tr>
<td>but</td>
<td>p_hyphen</td>
</tr>
<tr>
<td>by</td>
<td>p_period</td>
</tr>
<tr>
<td>for</td>
<td>p_ques</td>
</tr>
<tr>
<td>from</td>
<td>p_quote</td>
</tr>
<tr>
<td>had</td>
<td>p_semi</td>
</tr>
<tr>
<td>have</td>
<td>said</td>
</tr>
<tr>
<td>he</td>
<td>she</td>
</tr>
<tr>
<td>her</td>
<td>so</td>
</tr>
<tr>
<td>him</td>
<td>that</td>
</tr>
<tr>
<td>his</td>
<td>the</td>
</tr>
<tr>
<td>i</td>
<td>this</td>
</tr>
<tr>
<td>in</td>
<td>to</td>
</tr>
<tr>
<td>is</td>
<td>was</td>
</tr>
<tr>
<td>it</td>
<td>which</td>
</tr>
<tr>
<td>me</td>
<td>with</td>
</tr>
<tr>
<td>my</td>
<td>you</td>
</tr>
</tbody>
</table>
Hierarchical clustering

Allison et al. 2009
Allison et al. 2009
“But there is also a simpler explanation: namely, that these features which are so effective at differentiating genres, and so entwined with their overall texture – these features cannot offer new insights into structure, because they aren't independent traits, but mere consequences of higher-order choices. Do you want to write a story where each and every room may be full of surprises? Then locative prepositions, articles and verbs in the past tense are bound to follow. They are the effects of the chosen narrative structure.”
Project presentation

Tuesday April 25 (3) + Thursday April 27 (3)

12 min presentation +
5 min questions
YOUR CONFERENCE PRESENTATION

HOW YOU PLANNED IT:

INTRODUCE YOURSELF  DESCRIBE OUTLINE OF TALK  MOTIVATION

START

APPLAUSE

ENGAGING Q&A

RESULTS

CONCLUSIONS

15 MINUTES

METHODODOLOGY AND EXPERIMENT DESIGN

HOW IT GOES:

PREVIOUS SPEAKER RUNS LATE AND EATS INTO YOUR TIME.

TECHNICAL DIFFICULTIES CONNECTING YOUR LAPTOP.

FORGET INTRODUCING YOURSELF.

ANNOYING AUDIENCE MEMBER INTERRUPTS WITH SELF-AGGRAVATING QUESTION.

POWER THROUGH THE REST OF YOUR 30 SLIDES.

REALIZE YOU ONLY HAVE 3 MINUTES LEFT.

START

SPEND WAAAY TOO MUCH TIME DESCRIBING YOUR OUTLINE.

MOTIVATION

15 MINUTES

AWKWARD SILENCE Q&A.
Final report

• 8 pages, single spaced.

• Complete description of work undertaken
  • Data collection
  • Methods
  • Experimental details
  • Comparison with past work
  • Analysis

• See many of the papers we’ve read this semester for examples.
Final report

• **Clarity.** For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

• **Originality.** How original is the approach or problem presented in this paper? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

• **Soundness.** Is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments, proofs, or other argumentation?

• **Substance.** Does this paper have enough substance, or would it benefit from more ideas or results? Do the authors identify potential limitations of their work?

• **Evaluation.** To what extent has the application or tool been tested and evaluated? Does this paper present a compelling argument for

• **Meaningful comparison.** Do the authors make clear where the presented system sits with respect to existing literature? Are the references adequate? Are the benefits of the system/application well-supported and are the limitations identified?

• **Impact.** How significant is the work described? Will novel aspects of the system result in other researchers adopting the approach in their own work?