Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 5: Clustering overview

Feb 3, 2016
Clustering

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

$X = \text{a set of skyscrapers}$
Unsupervised Learning

- Matrix completion (e.g., user recommendations on Netflix, Amazon)

<table>
<thead>
<tr>
<th></th>
<th>Ann</th>
<th>Bob</th>
<th>Chris</th>
<th>David</th>
<th>Erik</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Star Wars</strong></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td><strong>Bridget Jones</strong></td>
<td></td>
<td>4</td>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><strong>Rocky</strong></td>
<td>3</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rambo</strong></td>
<td></td>
<td>?</td>
<td></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>task</td>
<td>$\mathbf{x}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learn patterns that define architectural styles</td>
<td>set of skyscrapers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learn patterns that define genre</td>
<td>set of books</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learn patterns that suggest “types” of customer behavior</td>
<td>customer data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods differ in the kind of structure learned
Hierarchical Clustering

• *Hierarchical* order among the elements being clustered
Dendrogram

Shakespeare’s plays

Witmore (2009)
http://winedarksea.org/?p=519
Bottom-up clustering

1. Given: a set $X = \{x_1, \ldots, x_n\}$ of objects
2. a function $\text{sim}: \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$
3. for $i := 1$ to $n$ do
4. \hspace{1em} $c_i := \{x_i\}$ end
5. $C := \{c_1, \ldots, c_n\}$
6. $j := n + 1$
7. while $C > 1$
8. \hspace{1em} $(c_{n_1}, c_{n_2}) := \arg\max_{(c_u, c_v) \in C \times C} \text{sim}(c_u, c_v)$
9. \hspace{1em} $c_j = c_{n_1} \cup c_{n_2}$
10. \hspace{1em} $C := C \setminus \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$
11. \hspace{1em} $j := j + 1$
Similarity

\[ \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R} \]

- What are you comparing?
- How do you quantify the similarity/difference of those things?
Probability

0.0  0.2  0.4

the  a  dog  cat  runs  to  store
Similarity

Euclidean = $\sqrt{\sum_{i}^{vocab} (P_{i}^{\text{Hamlet}} - P_{i}^{\text{Romeo}})^2}$

Cosine similarity, Jensen-Shannon divergence…
Cluster similarity
Cluster similarity

- Single link: two **most** similar elements
- Complete link: two **least** similar elements
- Group average: average of all members
Flat Clustering

- Partitions the data into a set of $K$ clusters
Flat Clustering

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

1. Given: a set $X = \{\tilde{x}_1, \ldots, \tilde{x}_n\} \subseteq \mathbb{R}^m$
2. a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$
3. a function for computing the mean $\mu : \mathcal{P}(<\mathbb{R}>R) \to \mathbb{R}^m$
4. Select $k$ initial centers $\tilde{f}_1, \ldots, \tilde{f}_k$
5. while stopping criterion is not true do
6.   for all clusters $c_j$ do
7.       $c_j = \{\tilde{x}_i \mid \forall \tilde{f}_l \ d(\tilde{x}_i, \tilde{f}_j) \leq d(\tilde{x}_i, \tilde{f}_l)\}$
8.   end
9.   for all means $\tilde{f}_j$ do
10.      $\tilde{f}_j = \mu(c_j)$
11. end
12. end
K-means
Representation

\[ x \in \mathbb{R}^F \]

[x is a data point characterized by F real numbers, one for each feature]

• This is a huge decision that impacts what you can learn
Voting behavior

<table>
<thead>
<tr>
<th>Issue</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes on abortion access</td>
<td>1</td>
</tr>
<tr>
<td>Yes on expanding gun rights</td>
<td>0</td>
</tr>
<tr>
<td>Yes on tax breaks</td>
<td>0</td>
</tr>
<tr>
<td>Yes on ACA</td>
<td>1</td>
</tr>
<tr>
<td>Yes on abolishing IRS</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ x \in \mathbb{R}^5 \]
First letter of last name

\[ x \in \mathbb{R}^{26} \]
## Representation

<table>
<thead>
<tr>
<th>task</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>learn patterns that define architectural styles</td>
<td>set of skyscrapers</td>
</tr>
<tr>
<td>learn patterns that define genre</td>
<td>set of books</td>
</tr>
<tr>
<td>learn patterns that suggest “types” of customer behavior</td>
<td>customer data</td>
</tr>
</tbody>
</table>
Evaluation

• Much more complex than supervised learning since there’s often no notion of “truth”
Internal criteria

• Elements within clusters should be more similar to each other

• Elements in different clusters should be less similar to each other
External criteria

• How closely does your clustering reproduce another (“gold standard”) clustering?
Learned clusters

A

B

C

Comparison clusters

D

E

F

G

H

I
Evaluation: Purity

- Learned clusters (as learned by our algorithm)
- External clusters (from some external source)

\[ \begin{align*}
G &= \{g_1 \ldots g_k\} \\
C &= \{c_1 \ldots c_j\} \\
\text{Purity} &= \frac{1}{N} \sum_k \max_j |g_k \cap c_j|
\end{align*} \]
\[
\frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j|
\]
Learned (G)

\[ = \frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j| \]
Learned (G)

\[ \frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j| \]

External (C)
$$\frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j|$$
(1 + 1 + 2) / 7 = .57
Evaluation: Rand Index

Every pair of data points is either in the same external cluster, or it’s not. = binary classification
Rand Index

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>same cluster?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubio</td>
<td>Paul</td>
<td>1</td>
</tr>
<tr>
<td>Rubio</td>
<td>Cruz</td>
<td>1</td>
</tr>
<tr>
<td>Rubio</td>
<td>Trump</td>
<td>0</td>
</tr>
<tr>
<td>Rubio</td>
<td>Fiorina</td>
<td>0</td>
</tr>
<tr>
<td>Rubio</td>
<td>Clinton</td>
<td>0</td>
</tr>
<tr>
<td>Rubio</td>
<td>Sanders</td>
<td>0</td>
</tr>
<tr>
<td>Paul</td>
<td>Cruz</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>Trump</td>
<td>0</td>
</tr>
</tbody>
</table>
Rand Index

21 decisions

\( N(N - 1)/2 \)
Learned

Predicted ($\hat{y}$)

<table>
<thead>
<tr>
<th></th>
<th>same cluster</th>
<th>different cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>same cluster</td>
<td>green</td>
<td>grey</td>
</tr>
<tr>
<td>different cluster</td>
<td>grey</td>
<td>green</td>
</tr>
</tbody>
</table>

True ($y$)

External
From the confusion matrix, we can calculate standard measures from binary classification.

The Rand Index = accuracy

\[
\frac{(1 + 12)}{21} = 0.619
\]
Example

Clustering characters into distinct types
The Villain

• Does (agent): kill, hunt, severs, chokes

• Has done to them (patient): fights, defeats, refuses

• Is described as (attribute): evil, frustrated, lord
The Villain

- Is character in the movie “Star Wars”
- Science Fiction, Adventure, Space Opera, Fantasy, Family Film, Action
- Is played by David Prowse
  - Male
  - 42 years old in 1977
Task

Learning character types from textual descriptions of characters.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>42,306 movie plot summaries</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>15,099 English novels (1700-1899)</td>
<td>HathiTrust</td>
</tr>
</tbody>
</table>
Evaluation I: Names

- Gold clusters: characters with the same name (sequels, remakes)
- Noise: “street thug”
- 970 unique character names used twice in the data; n=2,666
Evaluation II: TV Tropes

• Gold clusters: manually clustered characters from www.tvtropes.com

  • “The Surfer Dude”
  • “Arrogant Kung-Fu Guy”
  • “Hardboiled Detective”
  • “The Klutz”
  • “The Valley GIrl”

• 72 character tropes containing 501 characters
Purity: Names

Persona Regression

Dirichlet Persona
Purity: TV Tropes

![Bar chart showing comparison between Persona Regression and Dirichlet Persona for different dimensions: 25x25, 25x50, 25x100, 50x25, 50x50, and 50x100]
## Evaluation

<table>
<thead>
<tr>
<th>task</th>
<th>(\mathbf{x})</th>
</tr>
</thead>
<tbody>
<tr>
<td>learn patterns that define architectural</td>
<td>set of skyscrapers</td>
</tr>
<tr>
<td>styles</td>
<td></td>
</tr>
<tr>
<td>learn patterns that define genre</td>
<td>set of books</td>
</tr>
<tr>
<td>learn patterns that suggest “types” of</td>
<td>customer data</td>
</tr>
<tr>
<td>customer behavior</td>
<td></td>
</tr>
</tbody>
</table>
Digital Humanities

• Marche (2012), Literature Is not Data: Against Digital Humanities

• Underwood (2015), Seven ways humanists are using computers to understand text.
These charts show movie character interactions. The horizontal axis is time. The vertical grouping of the lines indicates which characters are together at a given time.

Lord of the Rings

Star Wars (original trilogy)
Characteristic vocabulary

Characteristic words by William Wordsworth (in comparison to other contemporary poets) [Underwood 2015]
Finding and organizing texts

- e.g., finding all examples of a complex literary form (Haiku).
- Supplement traditional searches: book catalogues, search engines.
Modeling literary forms

• What features of a text are predictive of Haiku?
Modeling social boundaries

Predicting reviewed texts [Underwood and Sellers (2015)]
Unsupervised modeling
Homework 1
Part one (everyone): Design an *ideal* representation of Oscar nominees to enable good prediction/analysis.
Part IIa. Implementation option. Instantiate a subset of those features for all nominees from 1960-2015. Deliverable: 6 feature files we will use to make predictions from.
<table>
<thead>
<tr>
<th>feature name</th>
<th>feature value</th>
<th>nominee canonical id</th>
</tr>
</thead>
<tbody>
<tr>
<td>boxoffice</td>
<td>60700000</td>
<td>/wiki/127_Hours</td>
</tr>
<tr>
<td>boxoffice</td>
<td>1000000</td>
<td>/wiki/12_Angry_Men_(1957_film)</td>
</tr>
<tr>
<td>boxoffice</td>
<td>168800000</td>
<td>/wiki/12_Monkeys</td>
</tr>
<tr>
<td>boxoffice</td>
<td>187700000</td>
<td>/wiki/12_Years_a_Slave_(film)</td>
</tr>
<tr>
<td>boxoffice</td>
<td>190000000</td>
<td>/wiki/2001:<em>A_Space_Odyssey</em>(film)</td>
</tr>
<tr>
<td>boxoffice</td>
<td>60400000</td>
<td>/wiki/21_Grams</td>
</tr>
<tr>
<td>boxoffice</td>
<td>2250000</td>
<td>/wiki/42nd_Street_(film)</td>
</tr>
<tr>
<td>boxoffice</td>
<td>9300000</td>
<td>/wiki/45_Years</td>
</tr>
<tr>
<td>boxoffice</td>
<td>5000000</td>
<td>/wiki/49th_Parallel_(film)</td>
</tr>
</tbody>
</table>
Part IIb. Critical option. The prediction process here is conditioned on being the nominee. Lots of public critique of the Academy this year for nominating no minority actors.

First, how would you model the Academy’s (human) nomination process? How might this result in the underrepresentation of minorities?

Second, consider an algorithmic approach to nominee prediction. What are the ways in which a similar underrepresentation can occur? What are the risks of training a supervised model?

How does representation of data influence these processes?

Deliverable: 3 page essay (single-spaced)