Definitions: AI, ML, DS

Jordan Boyd-Graber
University of Maryland
JANUARY 17, 2018
Roadmap

- What is data science / artificial intelligence / machine learning?
- What is possible with these new tools and techniques?
- What are the challenges?
Data provides the foundation
Machine learning provides algorithms
Artificial intelligence defines problems
Nowhere near true AI
Data are everywhere.
<table>
<thead>
<tr>
<th>Movie</th>
<th>Rating</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ikiru</em> (1952)</td>
<td>UR</td>
<td>Foreign</td>
</tr>
<tr>
<td><em>Junebug</em> (2005)</td>
<td>R</td>
<td>Independent</td>
</tr>
<tr>
<td><em>La Cage aux Folles</em> (1979)</td>
<td>R</td>
<td>Comedy</td>
</tr>
<tr>
<td><em>The Life Aquatic with Steve Zissou</em> (2004)</td>
<td>R</td>
<td>Comedy</td>
</tr>
<tr>
<td><em>Lost in Translation</em> (2003)</td>
<td>R</td>
<td>Drama</td>
</tr>
<tr>
<td><em>Love and Death</em> (1975)</td>
<td>PG</td>
<td>Comedy</td>
</tr>
<tr>
<td><em>The Manchurian Candidate</em> (1962)</td>
<td>PG-13</td>
<td>Classics</td>
</tr>
<tr>
<td><em>Midnight Cowboy</em> (1969)</td>
<td>R</td>
<td>Classics</td>
</tr>
</tbody>
</table>

User ratings
### Cheese

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Weight</th>
<th>Unit Price</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabot Vermont Cheddar</td>
<td>0.5 lb</td>
<td></td>
<td>$7.99/lb</td>
<td>$4.07</td>
</tr>
</tbody>
</table>

### Dairy

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Price per Unit</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendship Lowfat Cottage Cheese</td>
<td>1</td>
<td>$2.89/ea</td>
<td>$2.89</td>
</tr>
<tr>
<td>Nature's Yoke Grade A Jumbo Brown Eggs</td>
<td>1</td>
<td>$1.49/ea</td>
<td>$1.49</td>
</tr>
<tr>
<td>Santa Barbara Hot Salsa, Fresh</td>
<td>1</td>
<td>$2.69/ea</td>
<td>$2.69</td>
</tr>
<tr>
<td>Stonyfield Farm Organic Lowfat Plain Yogurt</td>
<td>1</td>
<td>$3.59/ea</td>
<td>$3.59</td>
</tr>
</tbody>
</table>

### Fruit

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Weight</th>
<th>Unit Price</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anjou Pears</td>
<td>3</td>
<td>1.76 lb</td>
<td>$2.49/lb</td>
<td>$4.38</td>
</tr>
<tr>
<td>Cantaloupe</td>
<td>2</td>
<td></td>
<td>$2.00/ea</td>
<td>$4.00</td>
</tr>
</tbody>
</table>

### Grocery

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Price per Unit</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fantastic World Foods Organic Whole Wheat Couscous</td>
<td>1</td>
<td>$1.99/ea</td>
<td>$1.99</td>
</tr>
<tr>
<td>Garden of Eatin' Blue Corn Chips</td>
<td>1</td>
<td>$2.49/ea</td>
<td>$2.49</td>
</tr>
<tr>
<td>Goya Low Sodium Chickpeas</td>
<td>1</td>
<td>$0.89/ea</td>
<td>$0.89</td>
</tr>
<tr>
<td>Marcal 2-Ply Paper Towels, 90ct</td>
<td>2</td>
<td>$1.09/ea</td>
<td>$2.18 T</td>
</tr>
<tr>
<td>Muir Glen Organic Tomato Paste</td>
<td>1</td>
<td>$0.99/ea</td>
<td>$0.99</td>
</tr>
<tr>
<td>Starkist Solid White Albacore Tuna in Spring Water</td>
<td>1</td>
<td>$1.89/ea</td>
<td>$1.89</td>
</tr>
</tbody>
</table>

**Purchases**
Knowledge

Akaike information criterion

Akaike information criterion (AIC) is an estimation procedure to determine the number of parameters in a statistical model. It is calculated using the following formula:

\[ AIC = 2k - 2\ln(L) \]

where:
- \( k \) is the number of parameters in the statistical model,
- \( L \) is the maximum likelihood of the estimated statistical model.

The preferred model is the one with the lowest AIC value. The AIC methodology is grounded in the concept of entropy. The AIC is an operational way of choosing among a finite number of models.

Definition

AIC = 2k - 2\ln(L)

where:
- \( k \) is the number of parameters in the statistical model,
- \( L \) is the maximum likelihood of the estimated statistical model.

Contents

- Introduction
- Definition
- Calculation
- Interpretation
- Comparison with other criteria
- Applications
- Criticisms
- References

Calculation

AIC is calculated as follows:

\[ AIC = -2\ln(L) + 2k \]

where:
- \( L \) is the maximum likelihood of the estimated statistical model,
- \( k \) is the number of parameters in the statistical model.

Interpretation

The AIC is a measure of model fit. A lower AIC value indicates a better fit of the model to the data.

Comparison with other criteria

The AIC is related to other information criteria such as the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQC).

Applications

The AIC is used in various fields such as ecology, econometrics, and machine learning.

Criticisms

The AIC has some limitations, such as the assumption of normality and the independence of the model errors.

References

Further reading on the topic of Akaike information criterion can be found in various academic journals and books.
Neuroscience
Social networks

Figure 2: Link communities in US direct flight data detected by Online MMLC. Each segment is less than 500 miles resulting in regional groups. Node sizes on top represent bridgeness, and on the bottom, influence.

MMSB [4] is a Bayesian probabilistic model of network data. MMSB is a mixed membership model. It associates each unit of observation with multiple clusters rather than a single cluster via a membership probability vector. It assumes \( K \) communities and directed edges are placed independently between node pairs with probabilities determined only by the pair's community memberships. Given the communities, the generative process defines a multinomial distribution over each node's...
Finance

Dow Jones Composite Average (^DJA) - DJI

4,372.78 ↓ 20.68 (0.47%) 1:48 PM EDT

Enter name(s) or symbol(s) | GET CHART | COMPARE | EVENTS | TECHNICAL INDICATORS | CHART SETTINGS | RESET

© 2012 Yahoo Inc.

Volume: 571,700

FROM: Aug 28 2012 TO: Sep 4 2012 -0.97%

Data are input to algorithms
We can recreate human decisions

![Diagram showing the process of feature extraction, classification, and spam detection. Features include items like viagra, bitcoin, subscribe, algorithm, bishop, and meeting.]
We can recreate human decisions

viagra
bitcoin
subscribe
algorithm
bishop
meeting

It's SPAM!
Algorithms can solve tasks
Predictions: Technology changes but society doesn’t
Are we prepared for social changes?
Discussions of AI Need Realism

- You can’t solve a task without good data
- Good data often requires humans
Discussions of AI Need Realism

- You can’t solve a task without good data
- Good data often requires humans
- Unless task is super-straightforward, also need ML expertise

![Graph showing the relationship between Accuracy and Intelligibility](image-url)
Discussions of AI Need Realism

- You can’t solve a task without good data
- Good data often requires humans
- Unless task is super-straightforward, also need ML expertise
- Effective solutions often require infrastructure
Discussions of AI Need Realism

- You can’t solve a task without good data
- Good data often requires humans
- Unless task is super-straightforward, also need ML expertise
- Effective solutions often require infrastructure
  - Explanations
  - Retraining
  - Socio-technical interfaces
Discussions of AI Need Realism

- You can’t solve a task without good data
- Good data often requires humans
- Unless task is super-straightforward, also need ML expertise
- Effective solutions often require infrastructure
  - Explanations
  - Retraining
  - Socio-technical interfaces
- DS/ML/AI is hard (but rewarding!)