Online Data Mining

PRESENTED BY Edo Liberty  www.edoliberty.com

Copyright © 2015 Yahoo All rights reserved. No reproduction or distribution allowed without express written permission.
Data mining in the batch setting
Batch computational model

\[ \begin{array}{ccccccc}
1 & 7 & 8 & 1 & \cdot & \cdot & 0 & 1 & 7 & 7 \\
\end{array} \]

Input

Computation

\[ \begin{array}{ccccccc}
a & b & b & b & \cdot & \cdot & c & a & c & b \\
\end{array} \]

Output
Distributed storage

The World

Data → Data → Data → Data → Computation → Result
Distributed model (map/reduce, message passing, ...)

The World

Data + Compute
Data + Compute
Data + Compute
Data + Compute

Data + Compute
Data + Compute
Data + Compute
Data + Compute

Computation

Result

Distributed model (map/reduce, message passing, …)
Distributed model (indexes, tables, databases, …)
Big-data *meta* infographic
The streaming computational model

The World

Computation

Sketch

Query Algorithm

Query

Result

Result
The streaming computational model

\[ O(n) \text{ Items} \]

Iterator

\[ O(\text{polylog}(n)) \] Computation per item

Computation

\[ O(\text{polylog}(n)) \] Space

Result

Sketch
The distributed streaming computational model

The World

Compute + Sketch
Compute + Sketch
Compute + Sketch
Compute + Sketch

Merge

Result

Query Algorithm

Query

Result
The distributed streaming computational model

Sketches Library from Yahoo!
A Java software library of stochastic streaming algorithms

Overview  Download  GitHub  Comments
Sometimes, you must take action immediately.
Machine Learning

Classify email as “spam” or “not”

Must be competitive with the best linear classifier in hindsight.

Prediction, Learning, and Games, Cesa-Bianchi, Lugosi, 2006
Online Portfolio Management

- Manage your portfolio online
- Be competitive with best CRP model (constant rebalanced portfolio)

Elements of information theory, Cover, 1991
Efficient algorithms for universal portfolios, Kalai, Vempala, 2003
Efficient Algorithms for Online Game Playing and Universal Portfolio Management, Agarwal, Hazan, 2006
Yahoo Finance App

Same story line or not?

1) The answer depends on the future
2) We have to decide now…
Online Feature Engineering

Learning feature representation with k-means, Adam Coates and Andrew Ng

On the Applicability of Unsupervised Feature Learning for Object Recognition in RGB-D Data, Manuel Blum, Jost Tobias Springenberg, Ja Wulfing, and Martin Riedmiller
Online model

1

Computation

Output
Online model

Input

Computation

Output

1 7

a b
Online model

Input

1 7 8

Computation

a b b

Output
Online model

Input

Computation

Output
Ski Rental Problem Example
Ski Rental

Rent: x$ /day

Buy: 1000$
Ski Rental

\[ 70 \]

\[ \text{Computation} \]

\[ R \]

\[ 70 + 70 = 140 \]
Ski Rental

Computation

R  R

70 + 90 = 160
Ski Rental

\[
\begin{align*}
70 & \quad 90 & \quad 80 \\
\text{Computation} & \\
R & \quad R & \quad B
\end{align*}
\]

\[
70 + 90 + 1000 = 1160
\]
Ski Rental

Computation

70  90  80  70

R  R  B
Ski Rental

\[
\begin{align*}
70 & + 90 & + 80 & + 70 \\
\hline
310
\end{align*}
\]

Much less than 1160…
Ski Rental

Input

| 70 | 90 | 80 | 70 | \[ \cdots \] | \[ \cdots \] | 90 | 88 | 72 | 79 |

Output

Computation

$1000$

\[ R \quad R \quad R \quad R \quad R \quad \cdots \quad \cdots \quad B \quad \square \quad \square \quad \square \quad \square \]
Online Regression

Online Principal Components Analysis, Boutsidis, Garber, Karnin, Liberty 2014
Online PCA with Spectral Bounds, Karnin, Liberty, 2015
Regression
Regression

\[ x_1, x_2, x_3 \]

\[ y_1, y_2, y_3 \]
Regression

\[ x_1, x_2, x_3, y_1, y_2, y_3 \]

\[ \| x_2 - \Phi y_2 \| \]
Online Regression

\[ x_1 \]

\[ y_1 \]
Online Regression
Online Regression
Online Principal Component Analysis

Online Principal Components Analysis, Boutsidis, Garber, Karnin, Liberty 2014
Online PCA with Spectral Bounds, Karnin, Liberty, 2015
Principal Component Analysis
Principal Component Analysis

\[ \Phi \Phi^T x \]
Principal Component Analysis

\[ \Phi \Phi^T x \]

\[ \| x_i - \Phi^T \Phi x_i \| \]
Online Principal Component Analysis

Algorithm 1 Fixed Error: Conceptual Algorithm

**input:** $X$, $\Delta$

$U \leftarrow$ all zeros matrix

for $x_t \in X$ do

if $\|(I - UU^T)X_{1:t}\|^2 \geq \Delta$

Add the top left singular vector of $(I - UU^T)X_{1:t}$ to $U$

yield $y_t = U^T x_t$

end for

Online PCA with Spectral Bounds, Karnin, Liberty, 2015
Online PCA, a visual example

Online PCA with Spectral Bounds
Online PCA combined with online learning

- Saves on storage and computation
- Does not hurt classification accuracy
- Is not worse than offline PCA and learning

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Before PCA</th>
<th></th>
<th></th>
<th>After PCA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dimension</td>
<td>Accuracy</td>
<td>Dimension</td>
<td>Online</td>
<td>Offline</td>
<td></td>
</tr>
<tr>
<td>census</td>
<td>401</td>
<td>0.93</td>
<td>40</td>
<td>0.94</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>maptask</td>
<td>5944</td>
<td>0.78</td>
<td>20</td>
<td>0.75</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>nomao</td>
<td>174</td>
<td>0.58</td>
<td>5</td>
<td>0.59</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>rcv1</td>
<td>43001</td>
<td>0.86</td>
<td>500</td>
<td>0.88</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>letter</td>
<td>16</td>
<td>0.75</td>
<td>2</td>
<td>0.76</td>
<td>0.76</td>
<td></td>
</tr>
</tbody>
</table>
Online k-means clustering

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko, 2014
k-means clustering

Points $x_i \in \mathbb{R}^d$

Small $\|x_i - x_j\|$ indicates the two points are “similar”
k-means clustering

Clusters $S_j$ -> Points $x_i \in \mathbb{R}^d$

A cluster is a set of points
k-means clustering

Clusters $S_j$

Points $x_i \in \mathbb{R}^d$

Centers $c_j \in \mathbb{R}^d$

Each cluster has a cluster center
k-means clustering

Clusters $S_j$

Points $x_i \in \mathbb{R}^d$

Centers $c_j \in \mathbb{R}^d$

K-means objective

$$\sum_{j=1}^{k} \sum_{i \in S_j} \| x_i - c_j \|_2^2$$
Hand written letters example

http://en.wikipedia.org/wiki/MNIST_database

http://research.ics.aalto.fi/mi/software/ne/
News groups example

- Roughly 20,000 documents
- 20 topics:
  - Graphics
  - PC hardware
  - Baseball
  - For-sale
  - Politics
  - ...

http://qwone.com/~jason/20Newsgroups/

http://research.ics.aalto.fi/mi/software/ne/
Same story line or not?

1) The answer depends on the future
2) We have to decide now...
Online k-means clustering

1) One can clustering points (documents) fully online
2) Create only slightly more than k centers (story lines)
3) Be competitive with the best clustering to k clusters

<table>
<thead>
<tr>
<th>Algorithm 2 Online k-means algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>input:</strong> V, k</td>
</tr>
<tr>
<td>C ← first k + 1 distinct vectors in V; and n = k + 1</td>
</tr>
<tr>
<td>(For each of these yield itself as its center)</td>
</tr>
<tr>
<td>w* ← min_{v,v'∈C} |v - v'|^2/2</td>
</tr>
<tr>
<td>r ← 1; q_1 ← 0; f_1 = w*/k</td>
</tr>
<tr>
<td><strong>for</strong> v ∈ the remainder of V <strong>do</strong></td>
</tr>
<tr>
<td>n ← n + 1</td>
</tr>
<tr>
<td><strong>with probability</strong> p = min(D^2(v, C)/f_r, 1)</td>
</tr>
<tr>
<td>C ← C ∪ {v}; q_r ← q_r + 1</td>
</tr>
<tr>
<td><strong>if</strong> q_r ≥ 3k(1 + log(n)) <strong>then</strong></td>
</tr>
<tr>
<td>r ← r + 1; q_r ← 0; f_r ← 2 ⋅ f_{r-1}</td>
</tr>
<tr>
<td><strong>end if</strong></td>
</tr>
<tr>
<td><strong>yield:</strong> c = arg min_{c∈C} |v - c|^2</td>
</tr>
<tr>
<td><strong>end for</strong></td>
</tr>
</tbody>
</table>

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015
Online k-means clustering

Clustering online is competitive with batch k-means++

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015

k-means++: the advantages of careful seeding, Arthur, Vassilvitskii, 2006
Online k-means clustering

Clustering online allows for feature engineering for online machine learning!

<table>
<thead>
<tr>
<th>dataset</th>
<th># records</th>
<th>dimension</th>
<th>Before</th>
<th>After</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>letter</td>
<td>20000</td>
<td>15</td>
<td>0.7581</td>
<td>0.7653</td>
<td>0.94%</td>
</tr>
<tr>
<td>shuttle</td>
<td>43500</td>
<td>8</td>
<td>0.9247</td>
<td>0.9950</td>
<td>7.60%</td>
</tr>
<tr>
<td>skin</td>
<td>245057</td>
<td>2</td>
<td>0.9247</td>
<td>0.9957</td>
<td>7.67%</td>
</tr>
<tr>
<td>poker</td>
<td>946799</td>
<td>9</td>
<td>0.5436</td>
<td>0.6015</td>
<td>10.65%</td>
</tr>
</tbody>
</table>

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015
Rule of thumb: Sequential decision making in big data should be both streaming and online.

Thank you