Machine Learning
(theory and practice)

Charles Sutton
Introduction to Research in Data Science
University of Edinburgh
New methodology

• Analyzing computer programs
• Data mining
• Exploratory data analysis
• Home energy demand
• Computer security

New applications

• New model architectures
• Inference algorithms (e.g., high dimensional, streaming)
• Approximate learning methods

New methodology

New applications
Interactive machine learning

Data analysts are like cats.
1. Want to explore their data
2. Don’t know what they want.

Interactive machine learning for analysts
Whose information need is not explicit
Whose domain knowledge is difficult to encode
Allow analysts to explore intermediate results

… not just for dummies!
Per clustering accept / reject

Clustering: Partition of data

TINDER
Technique for INteractive Data Exploration via Rejection

[Srivastava, Zou, Sutton, ICML W/S 2016]
Association Rules

Database of transactions

Association rule mining

Find set of all rules

that have

Prob \( \text{Prob} \left( \begin{array}{c} \text{orange} \\ \text{grapes} \end{array} \right) \geq \alpha \)

Count \( \{ \text{orange}, \text{grapes} \} \geq M \)

Frequent itemsets

Count \( \{ \text{orange}, \text{grapes} \} \geq M \)

Why? Exploratory data analysis
Probabilistic Itemset Mining

Generative Model

To sample a transaction,

1. For each itemset, sample
   \[ z_S \sim \text{Bernoulli}(\pi_S). \]
2. Deterministically set
   \[ X = \bigcup_{z_s=1} S. \]

Inference

View pattern finding as set cover

Alternate set cover and parameter inference (structural EM)

Itemsets: E-step is submodular set cover

Sequences: Interleaving model for patterns with gaps

Patterns

\[
\begin{align*}
\{ \text{orange, grapes, apple} \} & \quad 0.0001 \\
\{ \text{orange, grapes} \} & \quad 0.02 \\
\{ \text{grapes} \} & \quad 0.5 \\
\end{align*}
\]

Transaction

[Fowkes and Sutton, KDD 2016, PKDD 2016]
# API Call Patterns: “Big Code”

Twitter4j Java Library

<table>
<thead>
<tr>
<th>ISM Variant</th>
<th>MAPO</th>
<th>UPMiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Fowkes &amp; Sutton, FSE ’16]</td>
<td>[Zhong et al, ’09]</td>
<td>[Wang et al, ’13]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Method</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>TwitterFactory.&lt;init&gt;</td>
<td>TwitterFactory.&lt;init&gt;</td>
<td>TwitterFactory.&lt;init&gt;</td>
</tr>
<tr>
<td>TwitterFactory.getInstance</td>
<td>TwitterFactory.getInstance</td>
<td>TwitterFactory.getInstance</td>
</tr>
<tr>
<td>Twitter.setOAuthConsumer</td>
<td>Status.getUser</td>
<td>TwitterFactory.&lt;init&gt;</td>
</tr>
<tr>
<td>Twitter.setOAuthAccessToken</td>
<td>Status.getText</td>
<td>TwitterFactory.getInstance</td>
</tr>
<tr>
<td>Status.getUser</td>
<td>ConfigurationBuilder.&lt;init&gt;</td>
<td>TwitterFactory.&lt;init&gt;</td>
</tr>
<tr>
<td>Status.getText</td>
<td>ConfigurationBuilder.build</td>
<td>TwitterFactory.getInstance</td>
</tr>
<tr>
<td>AccessToken.getToken</td>
<td>ConfigurationBuilder.&lt;init&gt;</td>
<td>TwitterFactory.&lt;init&gt;</td>
</tr>
<tr>
<td>AccessToken.getTokenSecret</td>
<td>ConfigurationBuilder.setOAuthConsumer</td>
<td>TwitterFactory.getInstance</td>
</tr>
<tr>
<td>ConfigurationBuilder.&lt;init&gt;</td>
<td>TwitterFactory.&lt;init&gt;</td>
<td>Status.getUser</td>
</tr>
<tr>
<td>ConfigurationBuilder.build</td>
<td>TwitterFactory.&lt;init&gt;</td>
<td>Status.getText</td>
</tr>
</tbody>
</table>

- **Two main types of Twitter initialization call**

  [Fowkes and Sutton, FSE 2016]
Syntactic Idioms in Code

... if (c != null) {
    try {
        if (c.moveToNext()) {
            number = c.getString(phoneColumn);
        }
    } finally {
        c.close();
    }
}
...

(a)

try {
    if ($(Cursor).moveToFirst()) {
        $BODY$
    }
} finally {
    $(Cursor).close();
}

(b)

IfStatement
  expression:
    c!=null
  then:Block
  body:IfStatement
    expr:MethodInvocation
      expr:var%android.database.Cursor%
      name:c
      name:moveToFirst
    then:Block
    number = c.getString(c.getColumnIndex(phoneColumn));
  finally:Block
    ExpressionStatement
      MethodInvocation
        expr:var%android.database.Cursor%
        name:c
        name:close

Allamanis and Sutton, FSE 2014
Example Idioms

From: Nonparametric Bayesian Tree Substitution Grammar

[Post and Gildea, 2009; Cohn et al, 2010]

channel=connection.
createChannel();

(a)

Elements $name=$(Element).
select($StringLit);

(b)

Transaction tx=ConnectionFactory.
getDatabase().beginTx();

(c)

catch (Exception e){
  $(Transaction).failure();
}

(d)

SearchSourceBuilder builder=
getQueryTranslator().build(
  $(ContentIndexQuery));

(e)

LocationManager $name =
(LocationManager) getSystemService( Context.LOCATION_SERVICE);

(f)

try{
  Node $name=$methodInvoc();
  $BODY$
}finally{
  $(RevWalk).release();
}

(g)

$BODY$

(h)

Transaction tx=ConnectionFactory.
getDatabase().beginTx();

(i)

LocationManager $name =
(LocationManager) getSystemService( Context.LOCATION_SERVICE);

(j)

try{
  Node $name=$methodInvoc();
  $BODY$
}finally{
  $(RevWalk).release();
}

(k)

Document doc=Jsoup.connect(URL).
userAgent("Mozilla").
header("Accept","text/html").
get();

(l)
Predicting Names of Methods

if (DEBUG) assert n >= 0;
int r = 0;
while (n >= MIN_MERGE) {
    r |= (n & 1);
    n >>= 1;
} return n + r;

convolutional attention mechanism

[Allamanis, Peng, and Sutton, ICML 2016]
• **Machine learning for software engineering**
  • ML / NLP for programming languages
  • Combining program analysis with probabilistic machine learning
  • Find patterns in program executions: debugging

• **Machine learning for data science**
  • Deep learning: Combining neural networks with prior knowledge
    • “interpretability bias”
  • Learning how to clean data
  • Interactive machine learning
  • Tools for monitoring models over time
  • Unsupervised and weakly supervised learning

• **Deep learning: Unsupervised, structured, transfer learning**
  • ML for computer security, NLP, sustainable energy…

CUP, Wed and Fri 4pm

[https://wiki.inf.ed.ac.uk/ANC/CharlesUncertainPeople](https://wiki.inf.ed.ac.uk/ANC/CharlesUncertainPeople)