
Martin Emms

March 25, 2014
Unsupervised Learning from Data

Machine Translation
Unsupervised Learning from Data

Machine Translation

Speech Recognition
Unsupervised Learning from Data

- Machine Translation
- Speech Recognition
- Topic Modelling
Unsupervised Learning from Data

- Machine Translation
- Speech Recognition
- Topic Modelling

Use Unsupervised Machine Learning
Unsupervised Learning from Data

Machine Translation
Speech Recognition
Topic Modelling

\[
\text{Use Unsupervised Machine Learning}
\]

Possibly other attempts to infer from Big Data will use \textbf{Unsupervised Machine Learning?}
In Machine Translation

Unsupervised methods are used to go from sentence pairs to word pairs.
In Machine Translation

Unsupervised methods are used to go from sentence pairs to word pairs

<table>
<thead>
<tr>
<th>la maison est grand</th>
<th>the house is big</th>
</tr>
</thead>
<tbody>
<tr>
<td>c'est un haricot vert</td>
<td>its a green bean</td>
</tr>
<tr>
<td>je le lui donne</td>
<td>I give it to him</td>
</tr>
</tbody>
</table>

⇒

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(\text{la}</td>
<td>\text{the})$</td>
</tr>
<tr>
<td>$p(\text{le}</td>
<td>\text{the})$</td>
</tr>
<tr>
<td>$p(\text{maison}</td>
<td>\text{house})$</td>
</tr>
<tr>
<td>$p(\text{bleu}</td>
<td>\text{blue})$</td>
</tr>
</tbody>
</table>
Done using EM Algorithm

- Incomplete data
  - if we had complete data, would could estimate model
  - if we had model, we could fill in the gaps in the data

- Expectation Maximization (EM) in a nutshell
  1. initialize model parameters (e.g. uniform)
  2. assign probabilities to the missing data
  3. estimate model parameters from completed data
  4. iterate steps 2–3 until convergence
Learning Word Translations

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
Learning Word Translations

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- After one iteration
- Alignments, e.g., between la and the are more likely
Learning Word Translations

• After another iteration
• It becomes apparent that alignments, e.g., between *fleur* and *flower* are more likely (pigeon hole principle)
Learning Word Translations

- la maison ... la maison bleu ... la fleur ...

- the house ... the blue house ... the flower ...

- eventually converges, inherent hidden structure found by EM
- **Learns** translation probabilities

\[
\begin{align*}
p(la|the) & \quad 0.453 \\
p(le|the) & \quad 0.334 \\
p(maison|house) & \quad 0.876 \\
p(bleu|blue) & \quad 0.563
\end{align*}
\]
Topics in Documents

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Documents exhibit multiple topics biology genetics computation...
Topics in Documents

- **topic** = distribution over words
- **document** = distribution over topics
Topics in Documents

can only directly see words

but can learn the rest by **unsupervised methods**
Example learned words-for-topics

**GENETICS**
- human
- genome
- dna
- genetic
- genes
- sequence
- gene
- molecular
- sequencing
- map
- information
- genetics
- mapping
- project
- sequences

**EVL. BIOL.**
- evolution
- evolutionary
- species
- organisms
- life
- origin
- biology
- groups
- phylogenetic
- living
- diversity
- group
- new
- two
- common

**COMPUTING**
- computer
- models
- information
- data
- computers
- system
- network
- systems
- model
- parallel
- methods
- networks
- software
- new
- simulations
Hidden Markov Models for Speech or Actions

- words/action visible as an *evidence* sequence
- can be modeled with a HMM with *transitions between states* and *visible evidence for a state*
Learning HMM from Data

Possible to use unsupervised learning to find probabilities concerning hidden variables from data with just visible evidence

Used in Speech Recognizers, Activity Recognizers