Machine Learning @ Amazon

Ralf Herbrich
Amazon
Background

1992 – 1997 (Berlin, Diploma)

1997 – 2000 (Berlin, PhD)

2000 – 2009 (Microsoft Research)

2009 – 2011 (Microsoft)

2011 – 2012 (Facebook)

2012 – Present (Amazon)
Overview

• What is Amazon?

• Machine Learning in Practise
  • Probabilities
  • Finite Resource

• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems

• Conclusions and Challenges
Our Customers
Amazon’s Virtuous Cycles

- Lower Cost Structure
- Lower Prices
- Selection & Convenience
- Sellers
- Growth
- Customer Experience
- Traffic
Overview

• What is Amazon?
• **Machine Learning in Practise**
  • Probabilities
  • Finite Resource
• **Machine Learning @ Amazon**
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Artificial Intelligence and Machine Learning

Science
- Computer Science
- Statistics
- Neuroscience
- Operations Research

Artificial Intelligence
- Knowledge Representation
- Knowledge Extraction
- Reasoning
- Planning

Machine Learning
- Rule Extraction from Past (Training)
- Forecast of Future (Prediction)
- Taking Actions Now (Decision Making)

2/21/17
Overview

• What is Amazon?
• Machine Learning in Practise
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Machine Learning: Formal Definition

• Labelled Data

\[ \{P(y|x, w)\}_{w \in \mathcal{W}} + \{(x_i, y_i)\}_{i=1}^{n} \mapsto (x \mapsto P(y|x)) \]

• Unlabelled Data

\[ \{P(x|z, w)\}_{w \in \mathcal{W}} + \{x_i\}_{i=1}^{n} \mapsto (x \mapsto P(z|x)) \]

• Probability is a central concept in Machine Learning!
Why Probability?

1. Mathematics of Uncertainty (Cox’ axioms)
Cox Axioms: Probabilities and Beliefs

- Design: System must assign degree of plausibility \( p(A) \) to each logical statement \( A \).

- Axiom:
  - \( p(A) \) is a real number
  - \( p(A) \) is independent of Boolean rewrite
  - \( p(A|C') > p(A|C) \) \( \land \) \( p(B|AC') = p(B|AC) \)
    \[ \Rightarrow p(AB|C') \geq P(AB|C) \]

**P must be a probability measure!**
Why Probability?

1. Mathematics of Uncertainty (Cox’ axioms)

2. Variables and Factors map to Memory & CPU
**Factor Graphs**

- **Definition:** Graphical representation of product structure of a function (Wiberg, 1996)
  - Nodes: □ = Factors  ○ = Variables
  - Edges: Dependencies of factors on variables.

- **Semantic:**
  
  \[
p(x) = \prod_f f(x_{V(f)})
  \]
  
  - Local variable dependency of factors

\[
p(a, b, c) = f_1(a) \cdot f_2(b) \cdot f_3(a, b, c)
\]
Inference in a Factor Graph
Factor Graphs and Cloud Computing

\[ p(\theta|X, Y) \propto \prod_i p(y_i|\theta, x_i) \cdot \prod_j p(\theta_j) \]

Belief Store ("Memory")

Message Passing ("Communicate")

Data Messages ("Compute")
Factor Graphs and MXNet

```python
In [118]:
import mxnet as mx

# A simple network
x = mx.sym.Variable('x')
y = mx.sym.Variable('y')
z = x * y
net = mx.sym.LinearRegressionOutput(z, name='squaredloss')
mx.viz.plot_network(net)
```

Out[118]:

![Factor Graphs and MXNet Diagram](image)
Why Probability?

1. Mathematics of Uncertainty (Cox’ axioms)

2. Variables and Factors map to Memory & CPU

3. Decouple Data Modeling and Decision Making
Infer-Predict-Decide Cycle

**Inference:**
\[ P(\text{Parameters}) + \text{Data} \rightarrow P(\text{Parameters} \mid \text{Data}) \]
- Requires a (structural) model \( P(\text{Data} \mid \text{Parameters}) \)
- Allows to incorporate prior information \( P(\text{Parameters} \mid \text{Data}) \)

**Prediction:**
\[ P(\text{Parameters}) + \text{Data} \rightarrow P(\text{Data}) \]
- Requires integration/summation of parameter uncertainty
- Does not change state!

**Decision Making:**
Loss(\text{Action, Data}) + P(\text{Data}) \rightarrow \text{Action}
- Business-loss not learning-loss!
- Often involves optimization!
Overview

• What is Amazon?
• Machine Learning in Practice
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Finite Resources: Time

• **Real-Time Prediction Service: 5ms**
  - Number of search & ads candidates to rank: 10,000 \(\Rightarrow\) 500 ns/candidate
  - Time to read from main memory: 10ns \(\Rightarrow\) 50 variables
  - **L1/L2/L3 ranking pipelines to use more complex predictors**

• **Real-Time Learning Service: 1B examples/day**
  - Number of second per day: 86,400 \(\Rightarrow\) 86,400 ns/example
  - **Time to write to RAM: 100ns \(\Rightarrow\) 864 writes only (excluding disc access!)**

• Sparse models are the result of time constraints!
Finite Resource: Cost

**Economics 101**

- Profit = Revenue − Cost
- In the long run, a business that generates negative profits is not viable!

<table>
<thead>
<tr>
<th>Facebook</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual Revenue</td>
<td>$17,928,000,000.00*</td>
</tr>
<tr>
<td>Daily Revenue</td>
<td>$49,117,808.22</td>
</tr>
<tr>
<td>Number of DAU</td>
<td>1,038,000,000**</td>
</tr>
<tr>
<td>Number of Story Candidates</td>
<td>1,500***</td>
</tr>
<tr>
<td>Number of Daily Stories</td>
<td>1.557E+12</td>
</tr>
<tr>
<td>Maximum Cost per Story Candidate</td>
<td>$0.0000315</td>
</tr>
</tbody>
</table>


It’s power, stupid!

Some constraints might not be obvious: building new datacenters and **powering** them is non-trivial.

Example: 1 GPU box = 20 CPU boxes (in terms of power consumption)
Overview

• What is Amazon?
• Machine Learning in Practise
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Locations

ML Seattle
ML Bangalore
ML Cambridge
ML Berlin
Ivona
ML Los Angeles
ML Bangalore
Machine Learning Opportunities @ Amazon

**Retail**
- Demand Forecasting
- Vendor Lead Time Prediction
- Pricing
- Packaging
- Substitute Prediction

**Customers**
- Product Recommendation
- Product Search
- Visual Search
- Product Ads
- Shopping Advice
- Customer Problem Detection

**Seller**
- Fraud Detection
- Predictive Help
- Seller Search & Crawling

**Catalog**
- Browse-Node Classification
- Meta-data validation
- Review Analysis
- Hazmat Prediction

**Digital**
- Named-Entity Extraction
- XRay
- Plagiarism Detection
- Echo Speech Recognition
- Knowledge Acquisition
Overview

• What is Amazon?
• Machine Learning in Practise
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Fashion Forecasting (2014 – 2016)

**Setting**

- Given past sales of a fashion product, predict market demand up to 18 months into the future

**Challenges**

- **Sparsity**: Huge skew – many products sell very few items
- **Size**: Variable for Amazon retail in buying but not a variable for end customers
- **Seasonal**: Most fashion products only run for one season – often 12 months in advance
- **Distributions**: Future is uncertain \(\Rightarrow\) predictions must be distributions
- **Scale**: 5M+ fashion products in each market \(\Rightarrow\) \(10^9\) training examples
- **Censored**: Past sales ≠ past demand (inventory constraint)
**Demand Forecasting**

*Example fashion product to illustrate the challenges of forecasting.*

**Training Range:** Non-fashion items have longer training ranges that we can leverage. Need to information share across new and old products.

**Seasonality:** This item has Christmas seasonality with higher growth over time. This is where we need growth features in addition to date features.

**Missing Features or Input:** Unexplained spikes in demand are likely caused by missing features or incomplete input data.
Learning and Prediction

\[ P(z_{it} | \theta) \sim \]

Learning

Forecasting

Model Parameters
Slow Moving Inventory

Typical midsize dataset:
• About 5M items
• About 4.5B item-days
• About 98% zero demand
Sampling Predictions

\[ P(z_{it} | \theta) \sim \]

- 0 or ≥1?
  Binary classification #1

- 1 or ≥2?
  Binary classification #2

- If ≥2:
  Count regression z-2
In Practice
Modelling Out of Stock

GLM

Bridge
Overview

• What is Amazon?

• Machine Learning in Practise
  • Probabilities
  • Finite Resource

• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems

• Conclusions and Challenges

1432 Girali (Grasping toy) - Selecta Wooden Toys/Selecta Spielzeug
by Selecta Spielzeug
⭐⭐⭐⭐⭐  3 customer reviews

Price: £12.03 & FREE Delivery in the UK. Details

Only 7 left in stock.
Sold by Alle-Spielwaren and Fulfilled by Amazon. Gift-wrap available.

Want it delivered to Germany by tomorrow, 18 March? Order within 5 hrs 41 mins and choose One-Day Delivery to Germany at checkout. Details

18 new from £7.11
- 10 cm / 4 in.
- This classic series of grasping toys has been perfected by Selecta for over 30 years.
- See more product details
Machine Translation Pipeline

Input Request → Input Normalization → Tokenization

Detection & Escaping of Non-translatables

Translation/Decoding → Lowercasing → Sentence Segmentation

Re-insertion of (converted) Nontranslatables → Recasing

Translated Request → Post-processing → De-Tokenization
Machine Translation: Deep Dive

\[ p(\text{English} \mid \text{German}) = \frac{p(\text{English}) \times p(\text{German} \mid \text{English})}{p(\text{German})} \]

\[ \propto p(\text{English}) \times p(\text{German} \mid \text{English}) \]

- **Language Model**: What are fluent English sentences?

- **Translation Model**: What English sentences account well for a given German sentence?
Overview

• What is Amazon?
• Machine Learning in Practise
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems
• Conclusions and Challenges
Automated Produce Inspection: The Goal

Current Inspection

<table>
<thead>
<tr>
<th>Defect</th>
<th>Decay</th>
<th>Bruising</th>
<th>Bruising</th>
<th>Overripe/Soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample #</td>
<td>Defect Cat.</td>
<td>Serious</td>
<td>Damage</td>
<td>Serious</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>0</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>1</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>240</td>
<td>9</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>% of Total</td>
<td>100%</td>
<td>4%</td>
<td>17%</td>
<td>7%</td>
</tr>
</tbody>
</table>

New Automated Inspection

Computer Vision

Correct recognition as “Okay” with 64.46%!
Challenges

• Illumination
• Clutter/Occlusions
• Viewpoint
• Scale
• Intra-class variability
Computer Vision Pipeline

- Segmentation
- Defect Mining
- Produce Classification
Predicting Longevity
Age Aligned Strawberries (Test Set)
Overview

• What is Amazon?
• Machine Learning in Practise
  • Probabilities
  • Finite Resource
• Machine Learning @ Amazon
  • Forecasting
  • Machine Translation
  • Visual Systems

• Conclusions and Challenges
Conclusions

• Machine Learning “translates” data from the past into accurate predictions about the future!

• In practice, probabilistic models and finite resources matter.

• Machine Learning helps to improve customer experience at Amazon!
Thanks!