Evaluating Machine Learned User Experiences

Asela Gunawardana
Intelligent User Experiences
Microsoft Research
The typical machine learning problem
Evaluation is easy: just measure $\sum_i l_i$ on the test set.
Thank You

Questions?
**Problem:** for real problems, we need to decide what labels $y_i$ to look at, and what loss function $L(\cdot,\cdot)$ to use.
But is this really a serious problem?

How hard can it be?

E.g. Netflix:

$$x_i = (\text{user}_i, \text{movie}_i)$$
$$y_i \in \{1,2,3,4,5\}$$
$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$
Fixing the labels and loss fixes the problem

The “Netflix problem” at NIPS is:

\[ U^T \times M \approx R \]
The user’s Netflix problem is:
Really?
Where are the stars?
A coder at a tech company wins a week-long retreat at the compound of his company's CEO, where he's tasked with testing a new artificial intelligence.

Starring
- Domhnall Gleeson
- Alicia Vikander
- Sonoya Mizuno
- Oscar Isaac
- Claire Seiby
- Symara Templeman
- Gana Bayarsaikhan

Director
- Alex Garland

Genres
- Dramas
- Thrillers
- Psychological Thrillers
- Sci-Fi Thrillers
- British Movies
- Sci-Fi & Fantasy
- Sci-Fi Dramas
Does our formulation of the problem really help users find things to watch?
Does predicting ratings help users find things to watch?
Predicting Ratings ≠ Predicting Usage

RMS Rating Error

Netflix

BookCrossing

Alg A

Alg B
Predicting Ratings ≠ Predicting Usage

Online Retail Purchases

![Graph showing the comparison between Alg. A and Alg. B in terms of Precision and Recall. The graph illustrates that Alg. A consistently outperforms Alg. B across different recall levels.]
Predicting Ratings ≠ Predicting Usage

![News Story Clicks Graph]

- **Precision** vs **Recall** for two algorithms: Alg. A and Alg. B.
Lesson:

The “standard,” “given,” or “commonly used” labels and loss functions may tell us very little about how useful the system is.
If not RMSE, what?
Precision/Recall?
Mean Avg Precision?
Precison @16?
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Not Recommended</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Not Recommended</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Problem:
False Positive/True Negative

Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Not Recommended</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Problem:
False Positive/True Negative
Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.
Problem #1

Our data isn’t an i.i.d. draw – it’s collected from a real running system.
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Not Recommended</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Problems:
- False Positive/True Negative
  - Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.
- True Positive
  - Maybe the user would have watched the video already, even if we didn’t predict it.
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Not Recommended</strong></td>
<td><strong>False Negative</strong></td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Problems:

- **False Positive/True Negative**
  
  Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.

- **True Positive**
  
  Maybe the user would have watched the video already, even if we didn’t predict it.
Problem #2

Measuring prediction accuracy doesn’t tell us how the system will influence user behavior.
Mesothelioma Diagnosis? - www.simonsonfirm.com/Mesothelioma
We are Committed to Helping You Get the Compensation You Deserve.
Why File a Lawsuit? - Mesothelioma Settlements - Where We Serve

Mesothelioma Diagnosis? - sokolovelaw.com
www.sokolovelaw.com/mesothelioma ▼ (888) 940-5538
You Didn't Deserve This Disease. Learn About Your Legal Options Now.
Services: Legal Consultation, Help For Caregivers, Help For Veterans, Mesotheliom...
“A+ Rating” – Better Business Bureau
Settlement Stories - Why Sokolove Law? - Free Consultation

Mesothelioma Diagnosis? - bergmanlegal.com
www.bergmanlegal.com/FreeEvaluation ▼
Don’t Be Fooled By A Claims Center. Receive Top Compensation. Call Now.
Rated WA Super Lawyers - Exclusively Mesothelioma - $500 M in Settlements
Settlement Stories - Free Claim Evaluation - Reviews
9121 2nd Ave #2100, Seattle, WA - Open today: 8:00 AM – 5:30 PM

Mesothelioma Cancer | Diagnosis, Treatment and Support
www.mesothelioma.com/mesothelioma/
Mesothelioma is an aggressive cancer affecting the membrane lining of the lungs and abdomen. Malignant mesothelioma is the most serious of all...
Mesothelioma Symptoms - Pleural Mesothelioma - Mesothelioma Causes - Biopsies

Mesothelioma - Overview of Malignant Mesothelioma Cancer
www.asbestos.com/mesothelioma/ ▼ Asbestos.com ▼
Nov 10, 2015 - Malignant mesothelioma is a rare, asbestos-related cancer that forms on the thin protective tissues that cover the lungs and abdomen.... Latency Period: It can take 20-50 years after asbestos exposure for mesothelioma to develop.... That's because this cancer can take anywhere from 20 to ...

Mesothelioma - Wikipedia, the free encyclopedia
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Asbestos - Mesotheliom - Peritoneal mesothelioma - Category:Mesothelioma

Mesothelioma
ABOUT SYMPTOMS TREATMENTS
A tumor of the tissue that lines the lungs, stomach, heart, and other organs.

Very rare
Fewer than 20,000 US cases per year
Requires medical diagnosis
Lab tests or imaging always required
Consult a doctor for medical advice
Sources: Mayo Clinic and others.

More about this condition
Mesothelioma Portal

Mesothelioma Claims - 1-800-713-6692
www.nationalmesotheliomaclaims.com
You don't have to sue anyone for financial compensation. The $30 Billion Asbestos Trust Fund was established to pay asbestos damage claims to mesothelioma patients and their...

Mesothelioma Cancer Book - 1-800-301-1845
www.mesothelioma-answer.org
This authoritative book, 101 Facts About Mesothelioma by Anna Kaplan, M.D. has received many positive reviews for providing clear and simple answers to the most common questions about mesothelioma cancer.

Mesothelioma Attorney Locator - 1-800-314-2423
www.mesothelioma-attorney-locators.com
This site allows you to easily find Mesothelioma attorneys located in your state. Patient advocates also help patients and their families understand what they need to know when selecting a law firm that will...

Mesothelioma Doctor Match - 1-888-886-4051
www.mesotheliomaDoctorMatch.com
Now you can find a top mesothelioma doctor quickly. They also help you to get an appointment more quickly while others may need to wait weeks or months to be seen. Often this can shorten the time before your treatment starts which...

Mesothelioma & Veterans - 1-800-726-7245
mesothelioma.veteransupport.us
Important information for Navy and other military vets who have been exposed to asbestos and now have health problems. Learn about options for making mesothelioma and asbestos cancer claims. Filing for veteran's benefits that can provide...

Official Mesothelioma Calculator - 1-800-818-7146
www.mesothelioma-case-calculator.com
The amount of compensation recovered in a mesothelioma case depends on what asbestos products you were exposed to, which state the case is filed in, and what law firm you hire. Some cases settle for...

Mesothelioma Survival Rate - 1-888-886-1830
www.mesotheliomasurvivalrate.com
How long do mesothelioma patients live after being diagnosed? What options are there for treatment? Can changes in diet help? This site answers these questions and others related to
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended</td>
<td>True Positive</td>
</tr>
<tr>
<td>Not Recommended</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

Problems:
- False Positive/True Negative
  - Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.
- True Positive
  - Maybe the user would have watched the video already, even if we didn’t predict it.
- False Negative
  - Maybe the user watched the video but hated it.
A better Netflix evaluation protocol

1. Log usage (not just ratings)
2. Train recommender on log data from before yesterday.
3. Recommend items for yesterday’s users.
4. Score against yesterday’s actual usage data:

<table>
<thead>
<tr>
<th></th>
<th>Actually Used</th>
<th>Actually Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Not Recommended</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Problems:

False Positive/True Negative
- Maybe the user didn’t know about the video, would have happily watched it if we actually recommended it.

True Positive
- Maybe the user would have watched the video already, even if we didn’t predict it.

False Negative
- Maybe the user watched the video but hated it.
Problem #3

The influence of our system may only manifest over the long term.
1. Our data isn’t an i.i.d. draw – it needs to be collected from a real running system.
2. Measuring prediction accuracy doesn’t tell us how the system will influence user behavior.
3. The influence of our system may only manifest over the long term.

How do we avoid being fooled about how useful our system is?
How not to be fooled

1. Identify what the goal is
   • Service usage
   • Sales
   • Ad monetization
   • User retention

2. Randomly assign users to a control and treatment group and measure improvement due to system, over time.

3. Use (with care) offline experiments to prioritize which experiments to run.
The objections

Experiments are expensive and time-consuming—can only try a handful of variations.

We can’t really expect scientists to build user-facing systems before they do science.

Besides, I’m are confident that <insert loss function here> will generally track <insert real criterion here>.

RMSE was good enough for Netflix: $1,000,000 says so.

The system owner is happy with improvements in my metric.
This is where you lost your wallet?

No, I lost it in the park. But this is where the light is.
Science is a bit like the joke about the drunk who is looking under a lamppost for a key that he has lost on the other side of the street, because that's where the light is. It has no other choice.

Noam Chomsky
(at least, according to the web)
Science is a bit like the joke about the drunk who is looking under a lamppost for a key that he has lost on the other side of the street, because that's where the light is. **It has no other choice.**

Noam Chomsky

(at least, according to the web)
Another choice: Build a new lamppost

(or at least a flashlight)

Joachims, KDD 2002→WSDM 2015: Use actual user behavior and mild assumptions about it to evaluate web search ranking.

Marlin et al, IJCAI 2011: How to estimate and account for selection bias in data sets.

Bottou et al, JMLR 2013: How to use data reweighting and a priori causal knowledge to correct for selection bias and make counter-factual inferences.

These issues have started to be addressed, and we need to more work that builds on this start.
Need data that
is collected through randomization of a real system
records what was presented to the user ("impression logs")
records why (inputs and sampling probability/density)
records what the user did