ACTIVE DISCOVERY IN BIG DATA

Roman Garnett
CSE 591

October 25, 2018
1. INTRODUCTION

Big Data and Active Learning
Big Data
Big data

Collecting massive amounts of data is becoming easier and commonplace.
Social networks

• Facebook has over 1.2 billion user accounts.
• Facebook stores over 30 petabytes of user data!

Image: Martin Grandjean / CC-BY-SA 3.0
Chemicals

- The ZINC database of purchasable compounds contains approximately 35 million entries.
- 10,000 new compounds every day!
Credit card transactions

- Over *26 billion* credit card transactions in the United States in 2012.
- Over *390 million* credit card accounts in the United States as of Q3 2013.

Image: Thomas Kohler / CC-BY-SA 2.0
• Collecting data is becoming easier and widespread.

• Analyzing these data, however, is often very expensive! (And isn’t getting any easier...)
Drug discovery

• Imagine having access to all 35 million purchasable compounds.
• Which of them show significant activity against a biological target?
• Even with high-throughput screening it would take a year to test them all!
Fraud detection

- Conducting a fraud investigation is *very expensive*, potentially requiring human experts.
- Even by temporarily shutting down a card, *we are losing potential sales!*

**Introduction**
Intelligence analysis

NSA Records Every Call Made in Unnamed Foreign Country

Documents leaked by former NSA contractor Edward Snowden reveal the agency makes a record of every telephone call in a specific, unnamed foreign country, and keeps the recordings for up to a month.

The National Security Agency reportedly has a system in place that makes a record of
Information overload
Making intelligent decisions

• Analyzing data is often very expensive!
• In such cases, we should think carefully about which data we analyze.
• Having a lot of data to choose from is both a blessing and a curse!
Active machine learning
Active learning: Example

• Imagine trying to learn to separate the blue points from the green points given examples.

• Given many examples, this is easy.

• What if we can only afford to choose a very small number of examples?

\[ p(y = \bullet \mid x, \mathcal{D}) \]
Active learning: Example

random sampling (accuracy: 90%)
Active learning

Image: Burr Settles
Active learning: Example

- Can we do better than random sampling?
- Idea: given a model

\[
p(y = \bullet \mid x, \mathcal{D}),
\]

choose the *most uncertain* point.
- Perhaps by focusing on the boundary, we can *learn faster*.
- This is known as *uncertainty sampling*, a simple example of active learning.
Active learning: Example

uncertainty sampling (accuracy: 95%)
Making intelligent decisions

• Active learning is a powerful and flexible paradigm.
• Traditionally, active learning has focused on predictive accuracy.
• There are many important real-world problems where this is not our main concern!
• A main focus of my research is making intelligent decisions when faced with expensive observations, whatever the goal might be.
2. ACTIVE SEARCH
Finding interesting points
Active search

• In *active search*, we consider active learning with an unusual goal: *locating as many members of a particular class as possible.*

• Numerous real-world examples:
  • drug discovery,
  • intelligence analysis,
  • product recommendation,
  • playing Battleship.

---

1Garnett, Krishnamurthy, Xiong, Schneider (CMU), Mann (Uppsal). ICML 2012.
Battleship!
Which is better?

This is a bit of an unusual setting—classification accuracy is not directly important!
Our approach

We approach this problem via *Bayesian decision theory*.

- We begin by defining a simple *utility function* naturally suited for this task.
- The location of the next evaluation will be chosen by *maximizing the expected utility*. 
The utility function

We begin by choosing the natural utility function for this problem, *the number of interesting points found among the observed points*. If the labels $y \in \{0, 1\}$, then given data $\mathcal{D}$,

$$ u(\mathcal{D}) \triangleq \sum_i y_i. $$
Expected utility: One-step lookahead

- Suppose we only have one evaluation remaining.
- We calculate the expected utility of choosing point $x^*$ from among the remaining points. This is easy.

$$
\mathbb{E}\left[u(x^*, y^*, D_{t-1}) \mid x^*, D_{t-1}\right] = u(D_{t-1})
+ 1 \times p(y^* = 1 \mid x^*, D_{t-1})
+ 0 \times p(y^* = 0 \mid x^*, D_{t-1})
= u(D_{t-1}) + p(y^* = 1 \mid x^*, D_{t-1}).
$$

- Therefore, our best choice is to simply select the point with the highest probability.
Multiple-step lookahead

• One-step lookahead is simple and fast, but it’s also *myopic*. Can we do better?
• What if we plan *even farther ahead*?
Multiple-step lookahead leads to nontrivial behavior

Unlike the simple greedy one-step lookahead policy, two- and more-step lookahead leads to nontrivial choices. Let $\delta \geq \varepsilon$, and consider two evaluations. Which point should we choose first?

\[ \begin{align*}
\text{one-step: } & \varepsilon + \delta \\
\text{two-step: } & 2\varepsilon + (1 - \varepsilon)\delta \\
\text{difference: } & \varepsilon(1 - \delta) > 0
\end{align*} \]

Choosing the low-probability node is always better!
In fact, we can extend this example in a surprising way!

- Looking farther ahead can always help by any arbitrary amount!
- Marginal gains are not always decreasing!
Lookahead can always help

**Theorem** (Garnett, et al.)

Let \( \ell, m \in \mathbb{N}^+ \), \( \ell < m \). For any \( q > 0 \), there exists a search problem \( \mathcal{P} \) such that

\[
\frac{\mathbb{E}_{\mathcal{D}}[u(D) \mid m, \mathcal{P}]}{\mathbb{E}_{\mathcal{D}}[u(D) \mid \ell, \mathcal{P}]} > q;
\]

that is, the \( m \)-step active-search policy can outperform the \( \ell \)-step policy by any arbitrary degree.
Expected utility: Two-step lookahead

- Suppose now we have two evaluations remaining.
- To calculate the expected utility of choosing a point $x^*$, we must marginalize the unknown label $y^*$ as well as the location of the final evaluation and its label. This is a bit harder.
Big ugly equation?

We have:

\[
\mathbb{E}[u(x^*, y^*, x_t, y_t, D_{t-2}) \mid x^*, D_{t-2}] = \int \int \int u(x^*, y^*, x_t, y_t, D_{t-2}) p(y^* \mid x^*, D_{t-2}) \times \\
\times p(x_t \mid D_{t-1}) p(y_t \mid x_t, D_{t-1}) \, dy^* \, dx_t \, dy_t
\]
Three- and more-step lookahead

• In general, finding the optimal choice in the $\ell$-step lookahead case may be performed recursively.

• However, naively, it requires marginalizing $\ell - 1$ unknown future observations, both their locations and associated labels. This is expensive. (Exponential in the number of points!)
CiteSeer data

- Includes papers from the 50 most popular venues present in the CiteSeer database.
- 42k nodes, 222k edges.
- We search for *NIPS* papers, 2.5k papers (6%).

![Diagram showing cites/cited by relationship between paper A and paper B](image)
Huge graph!
Cost of lookahead

<table>
<thead>
<tr>
<th>( \ell = 2 )</th>
<th>( \ell = 3 )</th>
<th>( \ell = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>166 s</td>
<td>( \approx 146 ) days</td>
<td>( \approx 30 ) 500 years</td>
</tr>
</tbody>
</table>

Lookahead can always help, but is it *hopeless*?
Theoretical results

- For *well-behaved* classifiers, the optimal point can’t be *too far* from the one with maximum probability!
- We can *derive bounds on expected utility* that we can use to *prune the search space*, dramatically reducing computation time and enabling farther lookahead.
Bounding expected utility

\[ \text{expected utility} \]

\[ \chi \]

- local upper bound
- global lower bound
Results: Speedup from pruning

<table>
<thead>
<tr>
<th></th>
<th>$\ell = 2$</th>
<th>$\ell = 3$</th>
<th>$\ell = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>no pruning</td>
<td>166 s</td>
<td>(\approx 146) days</td>
<td>(\approx 30,500) years</td>
</tr>
<tr>
<td>pruning</td>
<td>0.228 s</td>
<td>15.0 s</td>
<td>745 s</td>
</tr>
<tr>
<td>speedup</td>
<td>731</td>
<td>(8.42 \times 10^5)</td>
<td>(1.29 \times 10^9)</td>
</tr>
</tbody>
</table>
Experiment

• We select a single NIPS paper at random, and begin with that single positive observation.
• The one-, two-, and three-step lookahead approximations are applied for a given number of evaluations.
Results

The graph shows the number of targets found over the number of evaluations for different search algorithms:

- **3-step optimal**
- **2-step optimal**
- **1-step optimal**
- **Random**

The y-axis represents the number of targets found, while the x-axis represents the number of evaluations.
Three-step lookahead found 8.5% of the targets after scanning only 1.3% of the data, 6.5 times better than random search would have done.

Further experiments on several other datasets show similar results, sometimes with even more significant improvement from increasing the lookahead.
Active Search

- One can understand the increased performance from two- and more-step lookahead from a simple viewpoint.
  - Uncertainty sampling is pure *exploration*.
  - One-step lookahead is pure *exploitation*.
  - Two-step lookahead is the first point where exploration and exploitation are *simultaneously considered*.
- This behavior *automatically falls out* by choosing the correct utility function and applying Bayesian decision theory. No heuristics or tricks were required!
- We learned a lot from considering *lookahead*!
Drug discovery$^2$

- Goal: use multiple-step lookahead active search for improving *virtual screening*.

• Perhaps active search could be used for dealing with large-scale *combinatorial search spaces* for discovering *new materials*?

• Examples: novel *alloys* (high-entropy alloys, bulk metallic glasses), novel *catalysts*, etc.

• Key challenge: designing informative *classifiers*!
3. QUASARS

Cosmic lighthouses
Quasars are massive, incredibly bright, very distant objects. They are probably supermassive black holes at the cores of young, active galaxies.
Quasars are bright!

Seriously, quasars are very bright. They can be 100 trillion times brighter than the sun, or about 100 times brighter than the entire Milky Way galaxy.
Quasars are distant!

Quasars (thankfully!) are incredibly distant. They have redshifts from around $z = 0.06$ to $z > 7$, which implies they’re between hundreds of millions to tens of billions light years away.
Quasars are old!

Quasars are therefore *incredibly old*, giving us a glimpse into the nature of the early universe and galaxy formation.
What to quasars look like?

- Here we will consider *spectroscopic* measurements of quasars.
- In spectroscopy, we measure the *spectral flux* (emitted radiation per unit wavelength per area) over a range of wavelengths of light.
Emission lines

Average of the spectra for many quasars:

Spikes correspond to intra-atomic events at fixed energies! (Quantum mechanics to the rescue!)
Hydrogen emission lines, Lyman-α

![Graph showing transitions in the hydrogen spectrum](image)

- **Lyman series**
  - n = 1 to 2: 122 nm, 103 nm, 97 nm
  - n = 2 to 3: 95 nm, 94 nm
  - n = 3 to 4: 656 nm, 486 nm
  - n = 4 to 5: 434 nm
  - n = 5 to 6: 410 nm

- **Balmer series**
  - n = 2 to 3: 1875 nm
  - n = 3 to 4: 1282 nm
  - n = 4 to 5: 1094 nm

- **Paschen series**
  - n = 4 to 5: 1875 nm
  - n = 5 to 6: 1282 nm
  - n = 6 to 7: 1094 nm
What to quasars look like?

![Graph showing coadded flux $y(\lambda_{\text{obs}})$ vs. observed wavelength $\lambda_{\text{obs}}$ (Å). The flux is measured in $10^{-17}$ erg/s/cm$^2$/Å.](image)
The Lyman-\(\alpha\) forest
Damped Lyman-\(\alpha\) absorbers

- When a very large gas cloud (column density \(> 2 \times 10^{20} \text{ cm}^{-2}\)) intervenes the line of sight, it causes characteristic “damping wings” to appear in the absorption profile. These are called damped Lyman-\(\alpha\) absorbers (DLAs).
- DLAs are a direct probe of non-luminous neutral gas at densities close to those required to form stars.
- They provide a powerful independent check on models of galaxy formation in the early Universe (\(z \sim 2–5\)).
Damped Lyman-α absorbers
Damped Lyman-$\alpha$ absorbers

![Graph showing normalized flux vs. rest wavelength $\lambda_{\text{rest}}$ (Å).](image)
The state of the art
The model of visual inspection is *prone to errors* (tired grad students) and *inefficient*.

There’s just too much data to keep up with. The *Sloan Digital Sky Survey* (SDSS) has captured around 300,000 quasar spectra, and plans to measure *millions more* over the next few years.
Goal and approach (Garnett, et al. 2016)

- Goal: Put grad students *out of business*.
- Approach: Use tens of thousands of measured quasars to build a *probabilistic model* of quasar spectra, and use this to automatically infer whether there is a DLA in a given spectrum.
4. CONCLUSION
Conclusions

• When we have a lot of data but analysis is expensive, we should think carefully about our decisions!
• We should always focus on our goal, which is often not simply to learn the best model!
• We should be careful to find the data we actually want!
• As storage gets cheaper at a rate faster than analysis gets easier, this will only become increasingly important!