Lecture 1: Overview of Machine Learning

What is Machine Learning?

- Have/collection some “data”
- Analyze it to produce “knowledge/insight”
- Use that knowledge for some “task”

Traditionally this is partitioned into:
- supervised learning
- unsupervised learning
- reinforcement learning

But many novel forms exist and are discovered and invented and being developed.

Supervised Learning Applications

<table>
<thead>
<tr>
<th>Domain/problem</th>
<th>Response Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather</td>
<td>Temperature tomorrow</td>
</tr>
<tr>
<td>Health Informatics</td>
<td>Number of Patients with Influenza</td>
</tr>
<tr>
<td>Environmental Engineering</td>
<td>Predict soil contamination level</td>
</tr>
<tr>
<td>Commerce</td>
<td>Product demand</td>
</tr>
</tbody>
</table>

In many applications the response variables is numerical and not categorical. These are known as regression problems.

Unsupervised Learning

Clustering is often a form of data exploration allowing us to identify groupings that are otherwise not apparent.

<table>
<thead>
<tr>
<th>Domain/problem</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-array data</td>
<td>Similar activity patterns</td>
</tr>
<tr>
<td>Text</td>
<td>Word Classes</td>
</tr>
<tr>
<td>Customer Activity</td>
<td>Customer “types” (phone: web; movies; etc)</td>
</tr>
</tbody>
</table>
Unsupervised Learning

Association rules capture nuggets of imperfect prediction rules

<table>
<thead>
<tr>
<th>Domain/problem</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market-basket Records</td>
<td>Nappy &amp; Milk → Bread (10% / 80%)</td>
</tr>
<tr>
<td></td>
<td>Nappy &amp; Milk → Beer (10% / 75%)</td>
</tr>
<tr>
<td>Census Data</td>
<td>(Age &lt;16) → Not in Army (X% / 100%)</td>
</tr>
<tr>
<td></td>
<td>(Age &gt; 30) &amp; (Boston) → (has BA/BSc) (X% / Y%)</td>
</tr>
</tbody>
</table>

Reinforcement Learning

Agent can control environment (take actions) and gets occasional rewards. Goal is to maximize long term reward.

<table>
<thead>
<tr>
<th>Domain/problem</th>
<th>Aim of Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI Planning/ Robot Control</td>
<td>Plan course of actions to achieve explicit or implicit goal.</td>
</tr>
<tr>
<td>(Computer) Games</td>
<td>Auto-player</td>
</tr>
<tr>
<td>City Management</td>
<td>Traffic Control</td>
</tr>
<tr>
<td>Web site management</td>
<td>Contents/advertising policy</td>
</tr>
</tbody>
</table>

Novel Forms of ML

- Constrained Clustering
- Active Learning
- Collaborative Filtering
- Collective Classification
- Machine Translation: variable length sequences
- Many more...
- You can think about more forms and whether they are feasible

Supervised Learning

- We will focus for a while on supervised learning. Think about any application: e.g. character recognition.
  - Is this really possible?
  - When and why?
  - What would you expect of algorithm?
  - How?

Supervised Learning

- What does it look like?
- Raw data depends on the applications and our interpretation of it.
  - Let’s look at 4 cases:
    - Toy example: PlayTennis
    - Soybean data
    - Mutagenesis molecule data
    - Astrophysics time series data
  - Demo of data files

Data Representation

- Raw data depends on the applications and our interpretation of it.
Classifier

- Given data representation (e.g. as table of features/attributes)

- What a classifier does:
  - Input: new example
  - Output: new example's predicted label

- How should we represent the classifier?

Learning Algorithm

- How can we analyze the data to produce a classifier that produces correct predictions?

Evaluation

- How do we know that we got it right?

- Evaluating the Classifier (the output of the algorithm) for a specific data/application

- Evaluating the Learning Algorithm that should be effective for any/many applications.

Learning Algorithm

- In the next few lectures we will discuss 4 different algorithms:
  - Nearest Neighbors
  - Decision Trees
  - Linear classifiers
  - Bayesian Classifier