Project Guidelines
Projects!

• Goal: apply machine learning to an interesting task

• Proposal (due Oct 26\textsuperscript{th}): 1pg
  – Who is in your group
  – Your task (and why is it interesting?)
  – Where did/will you get your data?
  – Which ML algorithms will you try first?
<table>
<thead>
<tr>
<th>Task</th>
<th>Due Date</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposal (1 pg)</td>
<td>Due 11:59PM Tuesday, Oct 26</td>
<td>10 pts</td>
</tr>
<tr>
<td>Status Report (2 pg)</td>
<td>Due 11:59PM Tuesday, Nov 16</td>
<td>10 pts</td>
</tr>
<tr>
<td>Project Poster/Demo</td>
<td>Thursday, Dec 9</td>
<td>20 pts</td>
</tr>
<tr>
<td>Project Web page</td>
<td>Thursday, Dec 9</td>
<td>15 pts</td>
</tr>
</tbody>
</table>
Meetings

• Status discussion
  – Nov. 19

• Last-minute issues
  – Dec. 3

• Optional

• Sign-up procedure to appear on course page
How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the results
6) Probably: go to (1)
How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the results

6) Probably: go to (1)
What’s the right task (for the class)?

- **Okay**: choose interesting, standard ML data set from UCI repository
- **Better**: use pre-existing but unique/important data set (e.g. Netflix prize, Google n-grams)
- **Best**: choose novel, important task and gather new data
- **Project completion** is important
  - Choose something interesting, but also something you can get done!
- **Things to consider:**
  - Availability of data
  - “Munging” required
  - Your knowledge of the domain
Examples (1 of 5)

• Something from your research
• The $ ones:
  – Price prediction (e.g. stock market)
  – Box office success
  – The “next big sound” see: nextbigsound.com
  – Sports contests
• UCI Repository
  – Tons of tasks, wines, mushrooms, text...
Examples (2 of 5)

• More data sources
  – Data.gov – US State data (agriculture, spending, etc.), census data
    • Also: NYC Big Apps
  – Customer reviews (summarization, deception detection...)
    • Other item attributes from review?
  – Twitter
Examples (3 of 5)

• Some of my favorites:
  – Predicting blog “anger”
    • (I have a small data set for this)
  – Extracting events from newspapers
    • Part of a history project with people from UNL, UK
    • I have unlabeled text for this
  – Compressing the Google n-grams data set
    • Unprecedented coverage, but takes 150G
    • Could a good ML approximation be much smaller?
Examples (4 of 5)

- Generics in language

  Birds lay eggs
  Mosquitoes carry the West Nile Virus

  Horses are female
  Humans are seven feet tall

Can we build a predictor for this?
Examples (5 of 5)

• Auditing search engine bias
  – Sentiment detection works in the aggregate
  – Can we measure whether different engines favor particular viewpoints?

• Ranking CS PhD programs
  – Do a survey, build predictor of human rankings
Brainstorming project ideas

- What’s your *second* best project idea?