Project Guidelines

Projects!

- Goal: apply machine learning to an interesting task

- Proposal (due next week): 1 pg
  - Who is in your group
  - Your task (and why is it interesting?)
  - Where did/will you get your data?
  - What's your initial approach?
    - It's okay if you can't say much about algorithms yet

Important Rules of Thumb

- If possible – set aside test data now, don't examine until end of course
- Allow time for iteration
- Understand your results

Meetings

- Status discussion
  - May 22,23

- 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 algorithm
6) Analyze the results
7) *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 algorithm
6) Analyze the results
7) *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 algorithm
6) **Analyze the results**
7) *Probably: go to (1)*

What's the right task (for the class)?

- **Okay**: choose interesting, standard ML data set from UCI repository or similar
- **Better**: use pre-existing but unique/important data set
- **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 3)

- Something from your research
- The $ ones:
  - Price prediction (e.g. stock market)
  - Box office success
  - Sports contests

Examples (2 of 3)

- Data sources
  - Data.gov – US State data (agriculture, spending, etc.), census data
  - http://data.world
  - NYC Big Apps
  - City of Chicago data portal
  - www.kaggle.com
  - WikiData
  - Customer reviews (summarization, deception detection…)
  - Twitter API

Examples (3 of 3)

- Other things people have done:
  - Will you get into your target sorority? (based on income, hometown, major, activities, etc.)
  - SafeRide wait times
  - CTEC text -> score

Metrics

- Precision/Recall vs. Accuracy

- Important: Use the right metric

Peer Review!

- You will review ~3 other groups' project proposals and status reports
- Peer reviews are worth 5 points (the same amount as your project proposal and status report!)